The Persistence of Human Capital: Evidence from Two RCTs in Rural West Africa

Alex Eble, Ila Fazzio, Peter Boone, Baboucarr Bouy, Alpha Camara, Robin Lumsdaine, Pei-Tseng Jenny Hsieh, and Deoclecio Marinho*

September 2026

Preliminary and incomplete; please do not cite or circulate without permission

Abstract

Human capital is a key engine for the generation of prosperity and its transmission across generations. In pockets of extreme poverty in the developing world, even the building blocks of human capital, such as foundational literacy and numeracy, are often out of reach for most children. In this paper we study whether very large literacy and numeracy gains reaped in childhood in two such contexts persist, fade out, or grow when these children are seen in late adolescence. Two RCTs in West Africa showed that an intervention model combining teacher coaching and structured pedagogy raised literacy levels at the end of third grade in rural areas from zero to between 25 and 60%. We find that these original skill gains both grew over time and gave rise to accelerated acquisition of higher-level skills. At ages 16-18, we observe 60% to 84% literacy rates among study participants in intervention villages, compared to 26% to 28% in controls. Our ITT estimates indicate 0.80 to 1.40 SD higher overall test score differences, and large gains on 5th-10th grade reading and math skills not taught in the original interventions. We also observe positive impacts on aspirations and some evidence of impacts on gender norms. Our results show that even in extremely disadvantaged settings, sufficiently large investments in human capital can generate lasting benefits that compound over time.

*Eble (corresponding author): Columbia University and NBER; email: eble@tc.columbia.edu. Fazzio: Effective Intervention; email: if@effint.org. Boone: Effective Intervention and London School of Economics; email: pb@effint.org; Bouy: Ministry of Public Service, The Gambia; email withheld. Camara: UNICEF; email: aatamsir15@gmail.com. Lumsdaine: American University, Erasmus School of Economics, and NBER; email: robin.lumsdaine@american.edu. Hsieh: National Foundation for Educational Research; email: j.hsieh@nfer.ac.uk. Marinho: Independent Consultant; email: jayrson89@gmail.com. We acknowledge very helpful feedback from many seminar audiences, as well as Iggy Boggild-Jones and Wayne Sandholtz. We acknowledge financial support from the JPAL Learning for All Initiative, the Jacobs Foundation, and the Khalifa International Award for Early Learning. JEL codes: I25, O15. Keywords: Education, literacy, numeracy, poverty, aspirations

1 Introduction

The acquisition of productive skills is one of the great engines of human progress over the last few millennia. It expands human capacity for health, the generation of wealth, and the transmission of advantage across generations (Becker, 1994; Hanushek and Woessmann, 2008; Black and Devereux, 2011; Asher, Novosad and Rafkin, 2024). A bedrock of this process is the acquisition of foundational literacy and numeracy skills. Foundational literacy and numeracy—the ability to write and parse text and to perform basic mathematical calculations—enables a wide range of productive activity, spanning the use of medicines or written instructions to the acquisition of skills and knowledge which build upon these skills.

Despite a global increase in these skills over the last century, there are pockets of extreme poverty in the developing world where, absent external intervention, most children will reach adulthood without having acquired them. Until very recently, most interventions to teach these skills in such pockets of poverty have shown relatively small learning gains, meaning that we lacked the technology needed to close the gap even between rural areas and urban areas of very low income countries, to say nothing of the gap between these areas and wealthier countries. In the last ten years, however, a series of evaluations have shown that a "bundled" intervention, combining multiple interventions known to work in isolation, can raise learning in these settings far more than previously thought possible (Piper et al., 2018; Kerwin and Thornton, 2021; Gray-Lobe et al., 2022).

The most impactful instantiations of this model took place in rural areas of The Gambia and Guinea-Bissau, two very low income West African countries (Eble et al., 2021; Fazzio et al., 2021). They yielded the largest gains in foundational literacy and numeracy on record. After three years, children in villages assigned to receive the intervention scored 3.2 and 5.3 standard deviations better on Early Grade Reading and Math Assessment-style tests (i.e., EGRA and EGMA tests). Using a binary measure of literacy, at the end of the third year, when all study children were supposed to be in third grade, zero percent of children in control villages in either country reached an international standard of literacy; in the intervention villages, 25% of Gambian children and

60% of Bissau-Guinean children reached this standard.

This paper asks whether the largest of these gains reaped in childhood persist, fade-out, or grow throughout adolescence, and what broader effect they may have on the children who experienced them. We follow the participants from those two studies into late adolescence, collecting data on four families of outcomes: learning, schooling/time use, aspirations, and gender norms. For learning, we conduct detailed in-person assessments of foundational literacy and numeracy (covering the same competencies as the EGRA and EGMA tests in the original studies), higher-level reading and math skills expected of students in grades 5-8 according to the Global Proficiency Framework, and a set of non-cognitive skills pertinent to the later personal and professional success of the trial participants, agreed upon by the authors and a gathering of international and local experts. For the other three outcome families, we collect data via survey. We use simple ordinary least squares analysis, following the research design of the original studies, to estimate the long-term impact of the two interventions.

For learning, our central finding is persistence. Seven years after the last intervention class was delivered, we see retention of the original foundational literacy and numeracy skills the intervention targeted, as well as accelerated acquisition of a series of more advanced reading and math skills not taught by the original intervention. The consensus from the literature on early-childhood and early-grade interventions, drawn largely from wealthier countries, is that test score impacts fade: absent continued intervention, gains measured at the end of a program typically shrink toward zero within five to ten years (Protzko, 2015; Hart et al., 2024). The countervailing intuition is that skills beget skills – that a child who can read can use that skill to acquire more skill, and in using it, retain it. Our results are consistent with the latter.

In our primary outcome, a composite score of the reading and math assessments, we find that children in intervention villages score 0.80 SD (Gambia) and 1.40 SD (Guinea-Bissau) better than children in control villages. While the magnitude is smaller than before, we argue this is driven primarily by problems with using a test score standard deviation as a measure of learning to compare across time and space (Cascio and Staiger, 2012; Rossiter et al., 2026). More specifically, secular

learning gains in the control group over the past seven years have caused the SD measure to be far better-behaved than it was in the original studies, when the distribution of learning in control villages was compressed around zero. As a result, a one SD difference in the 2018 test reflects a much smaller gain in absolute skills than it does in the 2025 test (Eble and Escueta, 2023).

Breaking these tests into their component parts—measuring the foundational literacy and numeracy skills targeted by the intervention and, separately, higher-level skills not taught by the intervention—we see large learning gains in both. On the Part A assessments, which measure the same competencies, at the same level of difficulty, as the endline tests of the original studies, participants from intervention villages outscore their control counterparts by 25.7 percentage points in literacy and 13.6 percentage points in numeracy in The Gambia, and by 29.7 and 15.7 percentage points in Guinea-Bissau. These gaps persist despite a rising tide beneath them: control-group participants, essentially none of whom could read at the end of the original studies, have acquired substantial foundational skill in the intervening years, as they progressed through school and, we conjecture, as mobile phones made printed material a part of daily life in their villages.

The gaps also persist despite a mechanical force working against them: in the intervention group, the Part A distributions pile up against the top of the scale, and this ceiling likely truncates the differences we are able to measure. The gains from the original interventions were not a head start that others erased; they have endured – and, if anything, grown – through late adolescence. This is underscored by results using our same binary measure of literacy, fluent reading with comprehension. In this measure, we find that 26% of children in control villages in both countries are now literate. In intervention villages, however, 60% of children in The Gambia and 85% in Guinea-Bissau are literate. Using Oral Reading Fluency (ORF)—a measure capturing the number of words read correctly per minute with less concern about ceiling effects—we see a doubling of ORF from roughly 40-50 words per minute read in controls to 90-95 in intervention villages.

The Part B and Part C assessments reveal that children in intervention villages have learned higher-level skills at a much faster pace than their counterparts in control villages. These tests assess skills that anchor on skills from national and international curricula at grades 5 and 8,

respectively. On Part B, participants from intervention villages outscore their control counterparts by 14.7 percentage points in literacy and 15.4 percentage points in numeracy in The Gambia, and by 19.4 and 20.4 percentage points in Guinea-Bissau; because these skills are more difficult, this translates into a difference of between 50 and 90 percent more questions answered correctly on that part. Part C, which tests skills taught in high school, is much more difficult, and at least 70 percent of students in both arms in both countries get no questions correct. The control-intervention differences here, however reveal that the intervention also helped unleash right-tail talent in both reading and math. The average number of questions answered correctly is around 1 percent in control villages, except for literacy in The Gambia, where it is 5.8%. In intervention villages, this percentage is between twice as large (literacy part C in The Gambia) and nearly eight times as large (literacy part C in Guinea-Bissau). In short, the interventions appear to have put participants on a steeper trajectory, not merely a higher intercept.

At the end of these tests, we administered tests of a series of non-cognitive skills. The skills assessed were agreed by experts in The Gambia and international learning experts to be valuable for the economic futures of these individuals, covering three "competencies": problem solving; teamwork and collaboration; and critical and creative thinking. We see no evidence of an impact of the intervention on an overall measure of these skills or any of the three components in The Gambia; in Guinea-Bissau, we see some evidence of a positive impact overall, loading entirely on the problem solving component.

Our next family of outcomes captures schooling, time use, and participation in the labor market. The primary outcome for this family is the highest grade in school the participant has attended. Here, we also see large differences: in The Gambia, children in intervention villages have gone 1.2 grades further in school than those in control villages. In Guinea-Bissau, this difference is 2.9 grades. Part of the disparity between the two countries comes from the fact that upper secondary schools are often further away from homes in Guinea-Bissau than they are in The Gambia, where a high school is usually within an hour's commute from most homes. This is reflected in the fact that in Guinea-Bissau, intervention children are 15.2 percentage points more likely to move than

control children; in The Gambia, we see no difference in likelihood to move. This analysis also reveals that, largely due to the fact that the majority of participants are in school in both countries, only a few dozen individuals in either country have regular paid work consistent with attachment to the labor market.

Our third family of outcomes pertains to the aspirations of the participant for their future. The primary outcome is a composite of three individual aspirations questions related to their highest level of education, migration from their home village, and their future line of work. We see that aspirations are significantly higher in the index among participants in intervention villages, as compared to those in control villages, and this difference is similar in both countries. We see the largest differences in education and work; for migration, the control mean is high enough that there is less scope for adjustment.

Our final family of outcomes pertains to gender norms. Here, the primary outcome is a composite of three questions related to gender norms, in particular polygamy and gender roles for male and female children in the home. We find that the intervention changed gender norms in Guinea-Bissau, with participants in intervention villages expressing 0.25 SD more progressive gender norms than those in control villages. In The Gambia, however, we find no significant difference, and this appears to be driven by a secular trend towards more conservative views in the country over time.

This last family of outcomes also contains a series of results related to realized and planned family formation. As with our labor market attachment results, we find that only a few dozen participants in either country have married or begun their fertility. Among plans for marriage and fertility, however, we find some evidence that the intervention may have shifted desired age at first child (Gambia) and total desired fertility (Guinea-Bissau), though these differences do not survive corrections for multiple hypothesis testing.

We analyze heterogeneity in impacts on the primary outcomes across the dimensions of heterogeneity analyzed in the original studies: gender, wealth, region, distance to main road, and caregiver education. The only result that survives multiple hypothesis correction is that, in The

Gambia, the intervention has a larger impact for boys than for girls. As in much of the rest of the world, girls in The Gambia often stay in school for longer than boys, even in rural areas. As a result, our findings tell us that the intervention helps close the female-male gap in education (favoring females) increasingly found in developing countries (Grant and Behrman, 2010). We do not, however, see any evidence of attendant heterogeneity by gender in learning.

Our findings add to three active bodies of literature in applied microeconomics. First, we contribute new evidence to the economics of education in developing countries, documenting that large investments in early education can have long and lasting benefits. This body of work has documented hundreds of interventions that have tried to raise learning levels (Evans and Popova, 2016; Glewwe and Muralidharan, 2016). Few have studied longer-term learning impacts, and those that do, find little evidence of impact (Barrera-Osorio, de Barros and Filmer, 2024). Our findings, that the large learning gains persist in both contexts long after the initial interventions and that they have effects both on schooling, higher-level learning, and aspirations, demonstrate the potential to expand human capacity by continuing such advancements here and in other similarly needy contexts. It also helps dispel pessimism that most efforts in such contexts can yield, at best, only moderate and mostly short-term returns (Pritchett, 2013; Evans and Yuan, 2022*a*).

We also contribute to the rich global literature, focused primarily on high-income countries, on the persistence and fade out of learning gains. The vast majority of studies to date have shown that even relatively large learning gains have something like a "half-life", in which they decay over time in the absence of ongoing scaffolding (cf., Chetty, Friedman and Rockoff 2014; Protzko 2015; Hart et al. 2024). Our findings show that a sufficiently large investment in human capital in a very low-resource setting exhibits growth over time, instead of this type of fade-out.

Finally, our results contribute to the broader set of studies of long-term impacts of high-intensity interventions designed to dramatically improve livelihoods for low-income individuals and households. These studies are united by having a larger per-unit (per-child, per-person, or per-family) cost, and include studies of graduation programs for the ultra-poor, cash transfers, and high school scholarships (Banerjee et al., 2015; Barham, Macours and Maluccio, 2024; Duflo

et al., 2026). Our findings suggest that there is a similarly large positive and sustained impact of a similarly high-fidelity foundational literacy and numeracy intervention. While our study takes place a similar or greater number of years after the end of the intervention as these other studies, it is too early yet to measure outcomes on fertility and labor market attachment. The large impacts on learning combined with other findings from the literature suggest, however, that positive impacts are likely to appear later in these individuals' lives, and contribute to an even larger societal benefit than initially estimated. We anticipate returning to these in future work.

The rest of the paper proceeds as follows. In Section 2, we describe the contexts in which we work and the original studies. In Section 3, we describe our research design. Sections 4, 5, 6, and 7 report findings on the four families of outcomes, respectively: learning, schooling, aspirations, and gender norms. Section 8 reports our heterogeneity analysis, and Section 9 concludes.

2 Background and Setting

Over the past 30 years, the challenge facing most low-income countries has changed from increasing enrollment to ensuring that learning happens for children when in school. This gap between schooling and learning is well documented: across dozens of countries, the majority of children fail to acquire the basic literacy and numeracy expected of them at early grade levels (Pritchett, 2013; Bold et al., 2017; Kremer, Brannen and Glennerster, 2013). This then follows them into higher grades, making the gap between curricular expectations and actual learning levels increasingly harder to bridge with age and grade (Muralidharan, Singh and Ganimian, 2019; Piper and Dubeck, 2024).

This shortfall is largest in the remote, rural areas that are home to the poorest households (Glewwe, 2002; Chaudhury et al., 2006; Glewwe and Muralidharan, 2016). In pockets of extreme poverty like those we study, the situation is even more severe still: high reported enrollment coexists with near-universal illiteracy. The large majority of children in these areas reach adulthood unable to read a simple sentence or perform elementary arithmetic (Sprenger-Charolles, 2008; Boone et al., 2014).

These foundational skills pose a particularly high return on investment for both the countries

in question and the broader world. Not only are literacy and numeracy the foundation on which most later human capital is built (Becker, 1994), they are also tightly linked to both individual and national paths to growth and prosperity (Woessmann, 2016). In these settings, weak family demand does not appear to be the binding constraint. Even caregivers in these settings who hold high aspirations for their children, and who invest accordingly, cannot on their own convert that demand into skills when the local supply of quality schooling is absent (Eble and Escueta, 2023; Fruttero, Muller and Calvo-González, 2025). Because these skills play a big role in determining the wealth and health of future generations as well as the current one, this disadvantage stands to reinforce itself over time. Any discussion of inequality, therefore, should include discussion of how to address the very low levels of learning in these places.

Recognizing this, researchers and governments have tested a very large number of interventions aimed at raising learning in low-income countries. Several large systematic reviews and meta-analyses synthesizing the accumulated evidence show that most interventions evaluated to date move learning levels by modest amounts, on the order of 0.1 to 0.5 standard deviations, and often less (Kremer and Holla, 2009; Evans and Popova, 2016; Glewwe and Muralidharan, 2016; Evans and Yuan, 2022*a*). Furthermore, these learning gains tend to be smaller in the poorest and most remote settings; precisely the settings where deficits are greatest. Programs that produce gains in learning of the magnitude that would bring children in these areas close to their peers in functioning school systems remain rare, and we are left with a striking mismatch: the need for large learning gains is most acute precisely where the accumulated evidence on how to achieve them is most scarce.

2.1 The original studies

Here we briefly describe the context, interventions, and participants in the original studies, and provide a very brief overview of their results. The organization running these studies sought to find governments who were interested in partnering to adapt and evaluate a bundled intervention originally shown to be highly effective in India, shifting the focus to be exclusively on the foundational literacy and numeracy skills taught in national curricula in grades 1-3. (Lakshminarayana et al.,

2013). The original sites were selected with three simple criteria: they were to be rural areas of a very low-income country where there was evidence of extremely low learning levels.

In both cases, the original intervention was adapted to the local context with national curricular experts in each country. Two key features of that intervention which were maintained in the adaptation were structured pedagogy and teacher coaching.[1] Structured pedagogy aims to give teachers scaffolding for teaching effectively. This includes specific guidance on what needs to be taught in each small increment of time, and clear instructions on how the teaching should be conducted—for example, guidance on what exercise to put on the board, how many students to call on, and suggested language on how to correct errors, recognize strong performance, and instruct on the correct processes to reach an answer. A series of evaluations have shown that structured pedagogy implemented on its own can raise foundational literacy and numeracy levels in low-income settings (Evans and Popova, 2016; Piper and Dubeck, 2024).

Teacher coaching refers to the broad set of practices in which teacher monitoring efforts focus on professional development and improvement of teaching skill. The implementation of teacher coaching in these settings involved hiring highly effective and experienced teachers, training them in the curricular model being used, and then sending them to monitor the program's teachers multiple times per month with the goal of improving the teachers' practice. Each coaching visit involved the coach arriving at the school before the lesson started, staying for the full day of instruction, observing the lessons, modeling best practice where appropriate, conducting rapid assessments of a sample of students during the day, and then, at the end of the day, sitting with the teacher to review the lessons, provide constructive feedback on ways to improve practice, and work with the teacher to address any major issues. If needed, coaches were also able to reallocate resources and, when needed, teachers, across classrooms, so that teachers needing assistance could observe a highly effective teacher for a short period in order to improve their own practice. Meta-analysis shows that this approach regularly improves both teacher instructional practices and student achievement (Kraft, Blazar and Hogan, 2018).

[1]These features were not known by those names at the time of the evaluation of the Indian intervention, but the intervention and its adaptation hewed closely to their core principles all the same.

Our studies combined these two components, along with flexible contracts for teachers[2] and provided sufficient funding for the interventions to operate at a high level of fidelity. The goal of the studies was to see how large of an impact on learning could be realized in very low-income settings if we combine multiple interventions known to work in isolation and substantially relaxed the budget constraint facing most interventions of this type.

Gambia In The Gambia, the study evaluated an after-school version of this intervention model. The intervention was administered by para teachers, delivering after-school classes following the local curriculum for 12 hours per week, starting in first grade, and following children from the first grade until the end of the third. The intervention cost approximately $232 per child per year (in 2016 dollars). The evaluation was conducted as a cluster randomized controlled trial in 169 villages (grouped into 111 clusters based on geographic proximity) in the two central regions of the country. Children were enrolled between the ages of 6 and 8 and were eligible only if they had not yet been to formal schooling, and their caregiver intended to enroll them in the first grade, for the first time, in the year to come. The study enrolled 4,518 children. The study's primary endpoint was each child's scores on endline tests of early grade reading and math ability.

Guinea-Bissau In Guinea-Bissau, the study evaluated an intervention which delivered kindergarten plus the first three years of primary schooling in lieu of government schools in rural parts of the country (as opposed to after school, as in The Gambia). This intervention was similar to the intervention in The Gambia in its three main features: one, that it followed an adaptation of the same set of structured lessons; two, that it relied heavily on teacher coaching to help teachers teach effectively; and three, that it used a large amount of financial resources (roughly $425 per child per year) to ensure that the intervention was implemented at a high level of fidelity. The evaluation was also conducted as a cluster randomized controlled trial and took place in 49 villages in the Quinara and Tombali regions of the country. Children were enrolled between the ages of 5 and 7 and were eligible only if they had not yet been to formal schooling. The study enrolled 2,112 children. The

[2]Also shown to raise learning in multiple developing country contexts (Muralidharan and Sundararaman, 2013; McEwan et al., 2015).

study's primary endpoint was each child's scores on endline tests of early grade reading and math ability that mirrored those administered in The Gambia.

2.2 Findings and implications

In both studies, we found that the intervention raised learning levels far more than previously thought possible. In The Gambia, we found that the intervention raised end-line test scores by 3.2 standard deviations. In Guinea-Bissau, we found that the intervention raised end-line test scores by 5.3 standard deviations. In follow-up work, we show that the standard deviation is an unreliable measure of learning in cases like these, where the control group distribution is compressed against the y-axis (Eble and Escueta, 2023). Using international benchmarks of absolute skill acquisition instead of the relative learning comparison that the standard deviation yields, we show similarly large gains in literacy and numeracy. In literacy, for example, a common international benchmark is for the child to be able to read at least 45 words per minute in a passage comprised of words at the third grade level or below, and correctly answer at least 80% of a short set of comprehension questions related to the passage (Dubeck and Gove, 2015; Fazzio et al., 2021). Using this benchmark, we find zero percent of children in the control group in both countries would be considered literate at the end of the third year of the study, in which the majority in both settings were enrolled in the third grade. In intervention villages, however, 25% of Gambian children and 60% of Bissau-Guinean children meet the standard for literacy. Our studies, in concert, with other studies in Kenya and Uganda (Piper, Simmons Zuilkowski and Mugenda, 2014; Kerwin and Thornton, 2021), showed that combining these two features can yield learning gains greater than the effect sizes expected from the sum of their parts.

3 Research design

This study follows participants from these two RCTs roughly seven years after the last data point was collected in the original studies. It was designed to answer two core questions: first, whether or not the large learning gains measured at the end of each study persisted in the years that had elapsed since the end of the intervention (and when these children had reached late adolescence,

i.e., between age 16 and 19); and second, whether or not those early learning gains had any impact on three other families of outcomes: schooling and time use; aspirations; and gender norms/family formation.

3.1 Data collection

Data was collected via surveys and learning assessments. The surveys were administered to the trial participants and their main caregivers, either in person or via phone. The assessments were administered to the trial participants in person. In addition, smaller assessments were administered to the next-oldest and next-youngest siblings of trial participants, conditional on the sibling in question being found in the village at the time of the trial participant survey.

The outcome variables we targeted were pre-specified and registered on the AEA RCT registry.[3] We group our outcomes into four families of outcomes: learning, schooling (which includes time use and participation in the labor market), aspirations, and family formation and gender norms. We collected these data via surveys, administered either in-person or via phone, and an assessment administered to all trial participants we could locate in person.

The next two subsections describe the assessment and the surveys in turn. These describe the methods for data collection and the nature of the datapoints collected. We define and describe the construction of primary and secondary outcomes from these datapoints in the results section of each relevant family of outcomes

3.1.1 The assessment

Our learning assessment aimed to assess learning among trial participants on three dimensions. The first is the persistence or fadeout of the original foundational literacy and numeracy skills on which both interventions focused. The second dimension is more advanced reading and math skills, expected of children around roughly grades 5 and 8, respectively; these were also linked to internationally recognized competencies for children at those levels, according to the UNESCO's Global Proficiency Framework.[4] The third dimension is a set of non-cognitive skills covering three

[3] Link: https://www.socialscienceregistry.org/trials/16285

[4] Source: https://www.uis.unesco.org/en/global-proficiency-frameworks.

component parts: one, problem solving; two, teamwork and collaboration; and three, critical and creative thinking.

To develop these assessments, we partnered with the National Foundation for Educational Research (NFER), a non-profit entity in the UK specializing in learning assessments and analytics. With the NFER and national curriculum experts in The Gambia and Guinea-Bissau, we developed a test of these skills, drawing on previous test items and linking reading and math competencies tested to national curricula. After conception, the assessments were piloted in each country and subsequently refined based on desirable psychometric properties such as reliability and validity before being used. They were kept strictly confidential over the period of the study.

For each participant, the learning assessment proceeded as follows. Participants were first administered one-on-one reading and literacy assessments of basic literacy and numeracy, following the style of EGRA and EGMA tests. We refer to these as the "Part A" literacy and numeracy assessments. Participants were then given written tests of higher-level reading and math skills, with items linked to the skills indicated for fifth grade in the Global Proficiency Framework. We refer to these as "Part B" literacy and numeracy assessments. The Part B assessments were then graded. If students performed above a prespecified threshold on a Part B assessment (either literacy or math), they were given a second written test, at a more advanced level, assessing even higher-level reading and math skills, with items linked to the skills indicated for eighth grade in the Global Proficiency Framework. Finally, we administered a many-on-one oral assessment of a series of non-cognitive skills grouped around three competencies relevant to success in the contexts in which these participants are likely to grow up: a. problem solving; b. teamwork and collaboration; and c. critical and creative thinking. These topics and the test items within were agreed upon after consultation with local curricular experts in each country as well as psychometricians and subject experts at the National Foundation for Educational Research. We show a summary of these tests and their content in Table 1.

Table 1: Content of the Learning Assessment

Skill	Part	Content
Literacy	A	Foundational literacy: letters, oral reading fluency, reading comprehension
	B	Grades 5-6 from local curriculum + grade 5 global proficiency framework: higher-level reading comprehension
	C	Grades 9-10 from local curriculum + grade 8 global proficiency framework: more complex reading comprehension. *Note: only administered if at least eight questions answered correctly on literacy part B*
Numeracy	A	Foundational numeracy (number recognition, sequences, up to two-digit addition and subtraction)
	B	Grades 5-6 from local curriculum + grade 5 global proficiency framework: reading graphs, basic statistics
	C	Grades 9-10 from local curriculum + grade 8 global proficiency framework: more complex calculations and logical puzzles. *Note: only administered if at least seven questions answered correctly on numeracy part B*
Life/non-cognitive skills	—	Teamwork and collaboration; problem-solving; critical and creative thinking

Notes: this table describes each of the learning assessments used in our study.

3.1.2 The survey

The survey contained questions focusing on three remaining families of outcomes. The first family is schooling and time use. Survey questions in this family capture information about the highest level of schooling the trial participant attained, their current and recent enrollment in school, their time use, and information about participation in the labor market. The second family is aspirations. Survey questions in this family capture the trial participant's aspirations for future education, employment, and migration. The third family is gender. Survey questions in this family capture information about the trial participant's gender norms, as well as their intended (or, if applicable, realized) family formation and fertility.

3.2 Empirical specifications

Our analysis in this paper follows the approach from the original studies. We regress our outcome variable on an indicator for treatment status, on the variables used for stratification, and nothing else, as given here:

$$y_{ic} = \beta_0 + \beta_1 Treatment_c + \beta_2 X_c + \epsilon_{ic} \tag{1}$$

The outcome variable is measured for individual i in cluster c. Treatment is assigned at the cluster level, and X is the (set of) control variable(s) for each country, which consists of the variable(s) used to stratify randomization and nothing else.[5] We cluster our standard errors at the cluster level, i.e., the level of randomization.

Our Y variables are all measured at the trial participant level, as described above. For each family of variables, we specify one primary outcome, and a series of secondary outcomes as described in our previously referenced pre-analysis plan. We also present some exploratory analysis; in those

[5] In The Gambia, randomization was stratified by two factors: region (Lower River or North Bank) and whether the village's distance to the main road was above or below median. In Guinea-Bissau, randomization was stratified by whether or not the village was above or below the median on a stratification composite variable comprising a weighted average of several indicators: the village's distance to the nearest road, the highest grade taught by the local school (in the one case where the village did not have a school, we set this to zero), the number of households in the village, the proportion of mothers speaking Crioulo in the village, and the third quartile of mothers' educational attainment in the village.

cases, the analysis is described as such.

3.3 Proportion of original participants reached at follow-up

We attempted to reach every participant still alive at the endline test in each study. We made multiple visits to each village, and worked with coordinators in the village to ensure that the trial participants were present in the village on the day of the survey or assessment. In cases where the trial participant had moved within the country, we sent teams to administer the survey and, separately, the assessment, to meet the trial participants in centralized locations. For the survey, if the trial participant could not be met in person—for example, if they were not able to make it to the village on the scheduled day, or they were out of the country—we conducted the interview by phone. For the assessment, between five and ten percent of participants reached refused; the most commonly given reason for refusal was that the trial participant could not read or write, and therefore did not wish to be assessed.

We describe these in Table A1, separately across all combinations of country (Gambia, Guinea Bissau) and instrument (survey, learning assessment). We see greater follow-up in the treatment arm in three of the four country-by-instrument cells. These differences are relatively small in The Gambia, on the order of 3 or 4 percentage points. In Guinea-Bissau, they are larger, up to 10 percentage points. For the assessment, a good deal of this relates to refusal to take the test because of inability to read. To account for potential bias from differential attrition across arms, we calculate Lee bounds for the primary outcome in each of our four outcome families (Imbens and Manski, 2004; Lee, 2009). We present these in Table A2; in all cases where we find significant results, the Lee Bounds we estimate also do not include zero.

4 Impacts on learning

Our primary outcome in the learning family is the participant's performance on a composite score of performance on parts A-C of the reading and mathematics assessments. This score captures a weighted average of the percent of questions answered correctly on each part, for each of the two tests. It is generated by first averaging scores across each part, within a subject (reading or math),

and then calculating the average between the two subjects. This follows the way that the primary outcomes were calculated in the original two studies.

Our pre-specified secondary outcomes include performance on each part. Following prior work, we present several transformations of these part-level performance measures. For part A, we present the performance transformed into a binary variable capturing literacy, following the same scoring criteria used in the original studies, as well as Oral Reading Fluency, calculated as the number of correct words read per minute in the reading passage section. For each part, we also present zero scores, i.e., the percent of participants answering zero questions correctly. Our final pre-specified secondary outcome is performance on the non-cognitive life skills assessment, standardized to zero mean and one SD, at the country-specific control group mean and standard deviation, respectively.

4.1 Results

We present our learning results in four steps. We begin with the primary outcome, the composite score, shown graphically in Figure 1 and in regression form in Table 2. We then decompose this result into its literacy and numeracy components and, within each, into performance on the three parts of the assessment. Recall that the assessment was designed to answer two distinct questions: Part A, which matches the competencies and difficulty of the endline tests from the original studies, measures whether the original learning gains persisted; Parts B and C, which test material the interventions never taught, measure whether those gains in foundational literacy and numeracy gave rise to greater acquisition of higher-level skills. Third, we use a simple binary measure of literacy that links these results to those of the original studies. Finally, we describe our results for the non-cognitive skills assessment.

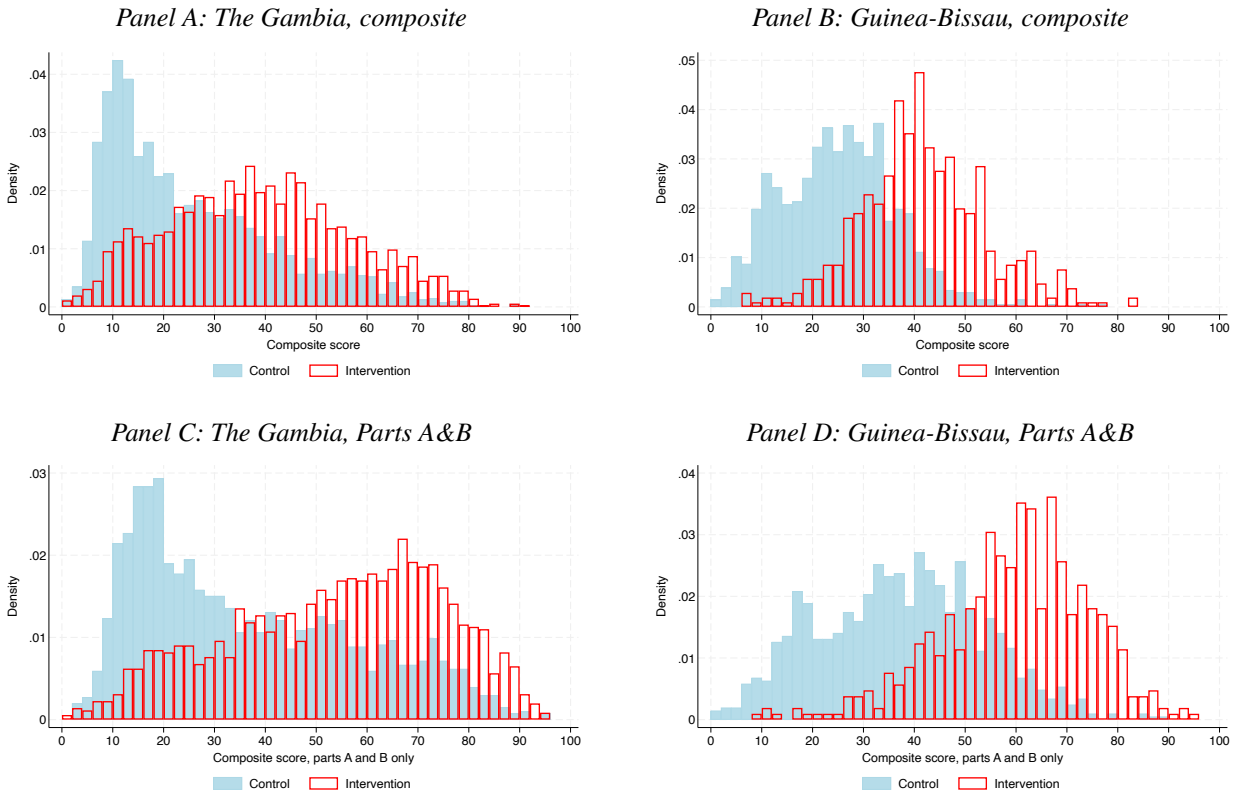
Seven years after our last contact with participants, we observe large differences in learning between individuals in intervention and control villages. Panels A and B of Figure 1 plot the distribution of the composite score, separately by treatment arm, for The Gambia and Guinea-Bissau, respectively, as indicated in the panel titles. In both countries, the distribution of scores in intervention villages sits well to the right of that in control villages. Table 2 presents the

corresponding estimates: participants in intervention villages score 13.4 percentage points higher than those in control villages in The Gambia (control mean: 25.6 percent) and 15.7 percentage points higher in Guinea-Bissau (control mean: 25.2 percent) – that is, they answer roughly 50 to 65 percent more questions correctly than a control-group baseline of about a quarter of questions answered correctly.[6] These estimates correspond to approximately 0.80 and 1.40 control-group standard deviations, respectively. Notably, these standardized effects are more readily interpreted than the 3.2 and 5.3 standard deviation estimates of the original studies: because learning levels in the control group have risen since endline, the control-group distribution is no longer compressed against zero, and a standard deviation now corresponds to a meaningful amount of absolute learning (Eble and Escueta, 2023). Because Part C is very difficult for all participants, the composite score compresses treatment-control differences; Panels C and D of Figure 1 therefore plot a composite of Parts A and B alone, where the differences are larger still, at 17.3 and 21.2 percentage points. Throughout this section and the rest of the paper, the estimates for Guinea-Bissau are larger than those for The Gambia, consistent with both the greater intensity of the intervention there—four years, delivered in lieu of school, in Guinea-Bissau, as compared with three years after school in The Gambia—and the weaker counterfactual schooling environment.

Figure 2 summarizes the Part A foundational literacy results with a single binary measure – the ability to read a simple printed paragraph and answer questions about its content. We plot this for each country at three time points: at baseline, where no children had this ability, at endline of the original studies, when most children were in third grade, and again in 2025. In 2018 at the original endlines, essentially no control participants in either country could do this, while roughly 25 percent (The Gambia) and 60 percent (Guinea-Bissau) of intervention participants could. Seven years later, roughly a quarter of control participants in both countries have reached literacy – the rising tide noted above – but the share of literate participants in intervention villages has continued to climb, to roughly 60 percent in The Gambia and 85 percent in Guinea-Bissau. The treatment-control gap in the probability of being literate is thus as large as, or larger than, it was at endline: the interventions

[6]Final "mop-up" data collection for the learning assessment in Guinea-Bissau was still underway at the time of writing; estimates for that country may change slightly in later drafts.

Figure 1: Distribution of composite test scores at follow-up, by treatment arm

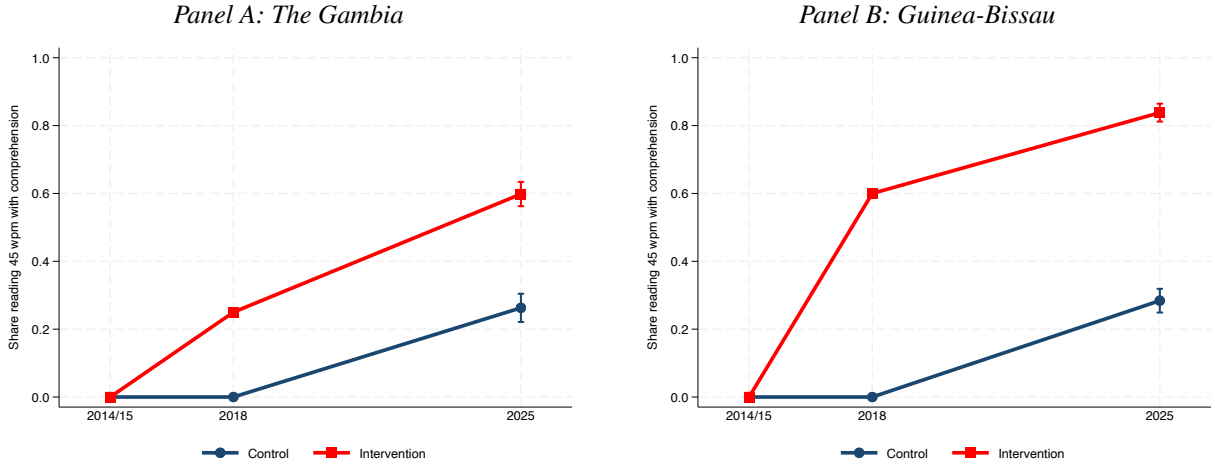


Notes: This figure plots the distribution of the composite test score at follow-up, separately for participants in intervention and control villages. Panels A and B show the composite of all three parts of the assessment; Panels C and D show the composite of Parts A and B only. Scores are expressed as the percent of questions answered correctly, weighted as described in the text.

appear to have put participants on a different learning trajectory, rather than merely giving a large early gain.

We can also use Item Response Theory to generate a measure of learning and use this as an outcome variable. We use part B and, separately, parts B and C together, to generate an item response theory measure of learning, and re-run our estimates. These appear in Table A3. They show standard deviation estimates very similar to the differences that we get from standard deviation measures of the tests, albeit with moderately smaller SD estimates: 0.67 SD in The Gambia, and 1.06 in Guinea Bissau.[7] We describe our implementation of this in Appendix A.

Figure 2: Share of participants who are literate at follow-up



Notes: This figure plots the share of participants meeting a binary threshold for literacy at follow-up, separately by country and treatment arm, with 95 percent confidence intervals.

We show distributions of test scores by treatment group for each of the three parts of the literacy exam in Figure 3. These show that the learning gains children in treatment villages have enjoyed in the intervening seven years extend well beyond what the original interventions taught. Both interventions ended after the third grade, and neither taught any of the material tested in Part B or C. Nonetheless, participants from intervention villages score substantially higher on both. In literacy, the treatment effect on Part B (material expected of fifth and sixth graders) is

[7]Following the original studies and guidance from our consultants at the National Foundation for Educational Research, we do not do this for Part A. We argue that the earlier, pre-literacy/numeracy decoding skills may constitute different dimensions of learning; the connected text measures in the EGRA test in particular have been shown to be unlikely to satisfy IRT models (Petscher, Mitchell and Foorman, 2015).

14.7 percentage points in The Gambia (control mean: 26.3 percent) and 19.4 percentage points in Guinea-Bissau (control mean: 15.4 percent). On Part C (material expected of ninth and tenth graders), control-group participants answer 5.8 percent of questions correctly in The Gambia and 0.9 percent in Guinea-Bissau; the treatment effects are 8.6 and 6.2 percentage points, respectively.[8]

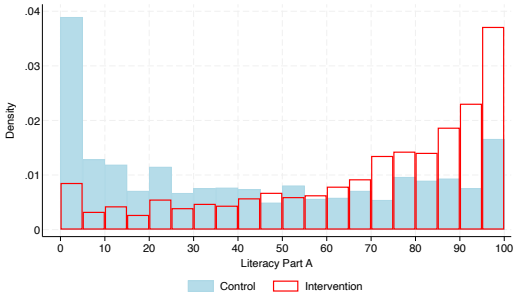
The numeracy results, appearing in Table 2 and Figures A3 and 4, follow the same pattern. Beyond the differences in means, the right tails of these distributions merit attention: the Part C distributions for intervention participants show a set of adolescents engaging successfully with genuinely difficult material – material that stumps nearly everyone in the control group. Statistically, there is no reason to believe the control villages hold fewer such talents; rather, the intervention appears to have given some participants the foundation to express them. While not definitive proof, this pattern is consistent with dynamic complementarity in skill formation: the foundational skills the interventions built appear to have raised the productivity of the schooling and other investments these participants encountered later. We are careful not to over-claim here – with these data alone we cannot yet separate how much of the gap reflects more years of schooling, more learning per year of schooling, or the direct compounding of early skills – and we return to formal tests of dynamic complementarity in ongoing work.

Another common transformation of performance on this kind of learning assessment is the zero score—a binary indicator for whether or not the test taker answered precisely zero questions correctly. In Appendix Table A4, we present results for zero scores on parts A, B, and C for literacy and numeracy separately for each country. These results show that the intervention significantly reduced the proportion of trial participants who answered no questions correctly on every skill-x-part cell, except for numeracy A, where essentially no students received a zero score. The largest of these gains are generally in Part C, where the treatment reduced the number of students who scored precisely zero on Part C by between 11.3 and 21.7 percentage points.

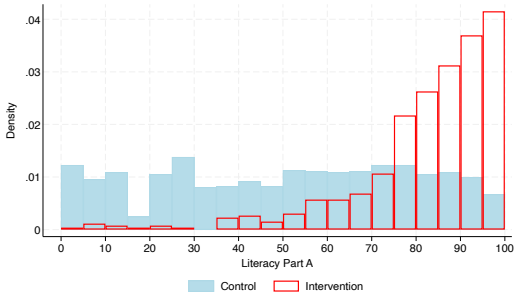
[8]Participants sat a Part C assessment only if they scored at least 8(7) of 25 literacy (numeracy) questions correctly on the corresponding Part B assessment; participants below this threshold are assigned a score of zero on Part C. Given the difficulty of Part C, we treat performance below this threshold on Part B as a sufficient statistic for a near-zero Part C score.

Figure 3: Literacy scores at follow-up by part, by treatment arm

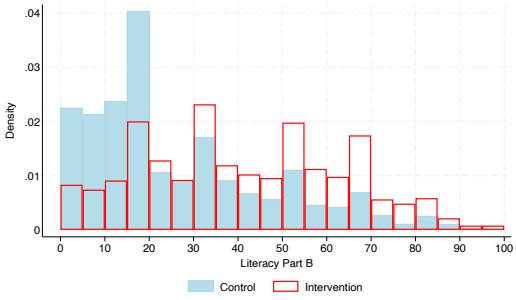
Panel A: The Gambia, Part A



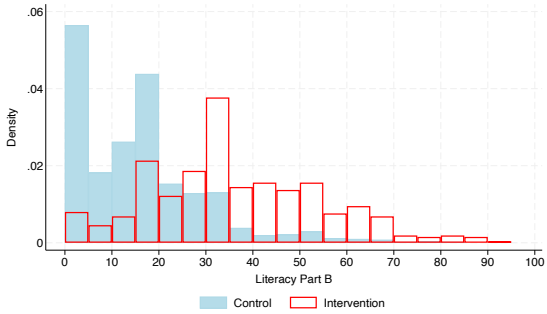
Panel B: Guinea-Bissau, Part A



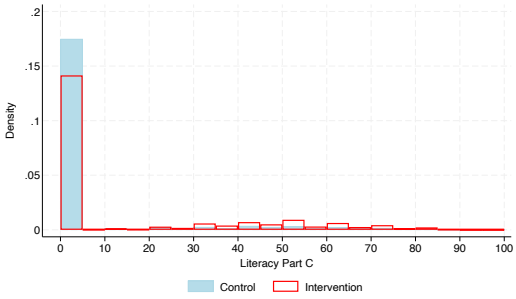
Panel C: The Gambia, Part B



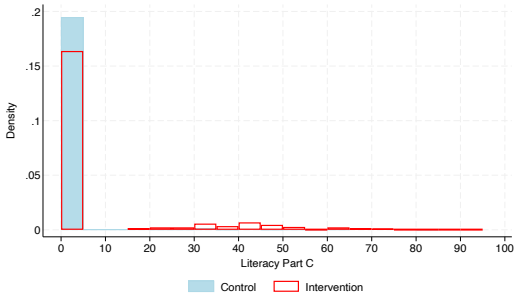
Panel D: Guinea-Bissau, Part B



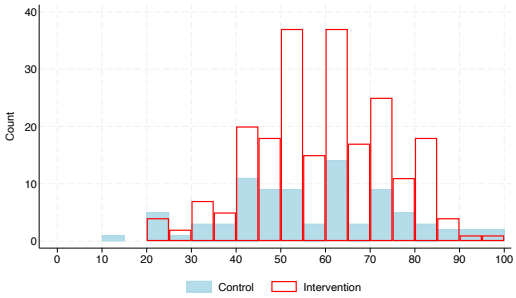
Panel E: The Gambia, Part C



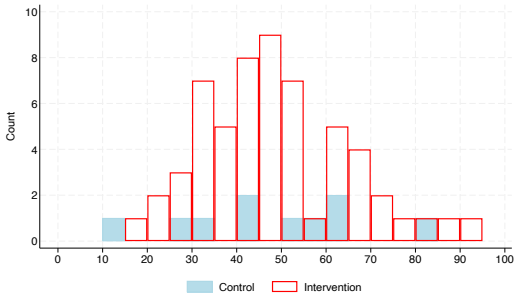
Panel F: Guinea-Bissau, Part C



Panel G: The Gambia, Part C nonzero scores



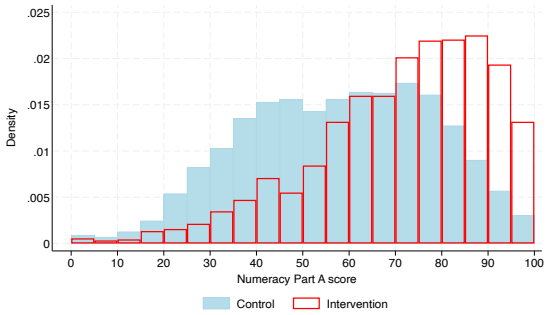
Panel H: Guinea-Bissau, Part C nonzero scores



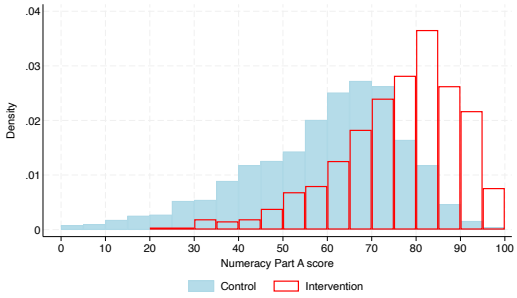
Notes: This figure plots the distribution of literacy scores at follow-up on each part of the assessment, separately for participants in intervention and control villages. Part A measures the foundational skills targeted by the original interventions (EGRA); Parts B and C measure higher-level skills linked to the Global Proficiency Framework for grades 5-6 and 9-10, respectively. In Panels G and H, we plot the frequency distribution of scores for only those who scored above zero on part C, to help with legibility of the graphs in panels E and F, given the large number of zero scores for part C.

Figure 4: Numeracy scores at follow-up by part, by treatment arm

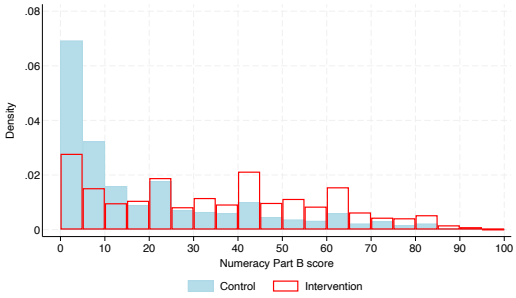
Panel A: The Gambia, Part A



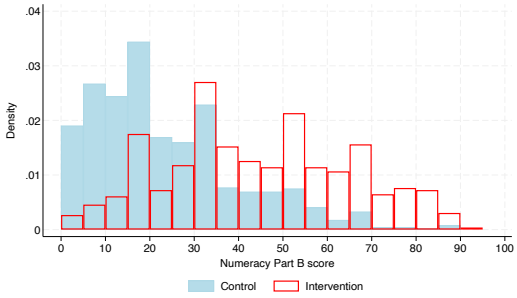
Panel B: Guinea-Bissau, Part A



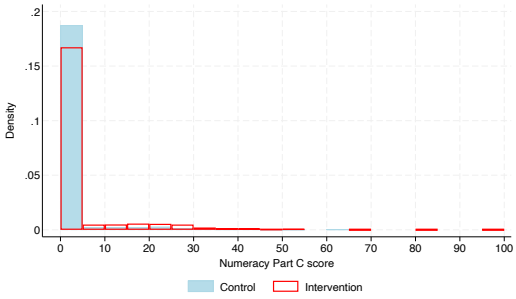
Panel C: The Gambia, Part B



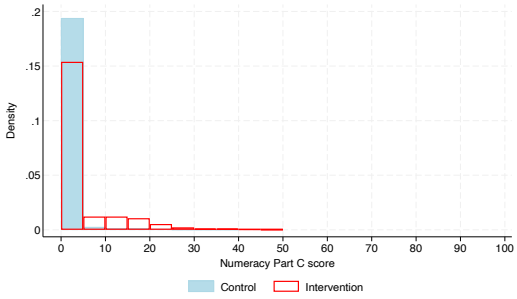
Panel D: Guinea-Bissau, Part B



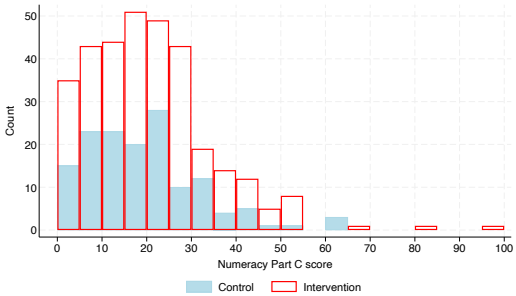
Panel E: The Gambia, Part C



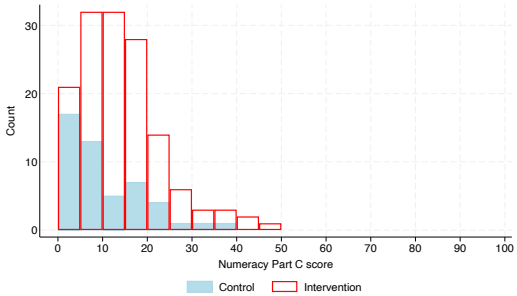
Panel F: Guinea-Bissau, Part C



Panel G: The Gambia, Part C nonzero scores



Panel H: Guinea-Bissau, Part C nonzero scores



Notes: this figure plots the distribution of numeracy scores at follow-up on each part of the assessment, separately for participants in intervention and control villages. Part A measures the foundational skills targeted by the original interventions (EGMA); Parts B and C measure higher-level skills linked to the Global Proficiency Framework for grades 5-6 and 9-10, respectively. In Panels G and H, we plot the frequency distribution of scores for only those who scored above zero on part C, to help with legibility of the graphs in panels E and F, given the large number of zero scores for part C.

Table 2: Treatment effects on learning at follow-up

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
			Composite scores			
Composite score	0.134***	0.256	3,799	0.157***	0.252	1,558
	(0.013)			(0.010)		
Composite score, parts A&B	0.173***	0.365	3,799	0.212***	0.374	1,558
	(0.017)			(0.012)		
			Literacy scores			
Literacy score (overall)	0.163***	0.249	3,802	0.184***	0.222	1,571
	(0.017)			(0.012)		
Literacy score, parts A&B	0.201***	0.345	3,802	0.246***	0.329	1,571
	(0.020)			(0.014)		
Literacy, Part A	0.257***	0.426	3,818	0.297***	0.504	1,571
	(0.024)			(0.019)		
Literacy, Part B	0.147***	0.263	3,825	0.194***	0.154	1,571
	(0.016)			(0.014)		
Literacy, Part C	0.086***	0.058	3,819	0.062***	0.009	1,571
	(0.012)			(0.012)		
Oral Reading Fluency	41.135***	48.134	3,736	52.432***	43.402	1,571
	(3.702)			(2.249)		
			Numeracy scores			
Numeracy score (overall)	0.104***	0.262	3,802	0.131***	0.279	1,558
	(0.011)			(0.009)		
Numeracy score, parts A&B	0.144***	0.385	3,802	0.180***	0.416	1,558
	(0.015)			(0.012)		
Numeracy, Part A	0.136***	0.578	3,818	0.157***	0.596	1,562
	(0.014)			(0.012)		
Numeracy, Part B	0.154***	0.192	3,825	0.204***	0.235	1,566
	(0.016)			(0.014)		
Numeracy, Part C	0.024***	0.014	3,818	0.034***	0.006	1,563
	(0.004)			(0.004)		

Notes: Each row reports the results of a separate regression of the indicated outcome on an indicator for assignment to the intervention and fixed effects for the randomization strata, following Equation 1. All scores are expressed as the share of questions answered correctly, except for ORF. Part A measures foundational literacy and numeracy (EGRA/EGMA); Parts B and C measure skills linked to the Global Proficiency Framework for grades 5-6 and 9-10, respectively. Participants sat Part C only if they scored at least 8 (7) of 25 literacy (numeracy) questions correctly on the corresponding Part B; participants below this threshold are assigned a Part C score of zero. *, **, and *** denote significance at the 10, 5, and 1 percent levels. Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum.

4.2 Non-cognitive skills

Our final analysis in this section reports the results for a set of non-cognitive skills, also known as life skills. We present the results first for an index of the three categories, and then the three categories separately: problem-solving and decision-making; critical and creative thinking; and teamwork and collaboration. The index equally weights each category. All scores are standardized to the country-specific control group mean and standard deviation.

Our results, given in Table A5, show no evidence of an impact of the intervention on the non-cognitive skills of treatment children in The Gambia. In Guinea-Bissau, we see a positive impact of the treatment on both the index and on the problem-solving and decision-making component. While the treatment effect estimates are not significant for the other two components, their magnitude is, in each case, more than twice that from The Gambia, suggestive of a change in these as well.

5 Schooling

The primary outcome in this family of variables is the highest grade of school the trial participant attended. Secondary outcomes include a set of outcomes capturing education and labor market attachment. For education, they include whether the trial participant is still in school at the time of survey and whether they were in school in the previous year. They also include the following outcomes, conditional on the participant being in school this year or the previous: annual educational expenditure on the trial participant's education, location of the school relative to trial participant's village of enumeration, and type of school. If the trial participant was not in school this year, secondary outcomes include the reason they left school, their intention to go back to school, their main daytime activity, whether they are employed, and, conditional on being employed, the type of job, hours worked per week, and weekly work income. For all trial participants, we also collected the highest grade attended of the participant's next older and younger siblings conditional on their being in the village at time of the survey.

5.1 Results

We begin with the primary outcome, the highest grade attended, shown in Figure 5 and in the first row of Table 3. We then discuss current enrollment and educational expenditure, time use, and migration and labor market outcomes.

In both countries, participants from intervention villages have reached higher grades than their control counterparts. In The Gambia, the impact of the treatment on highest grade reached is 1.19 grades (control mean for highest grade: 6.98). In Guinea-Bissau, the impact is 2.86 additional grades (control mean: 4.97 grades) – an increase of more than half of the control mean.

When we last saw these participants, at the endline of the original studies, there was little difference in grade attainment across arms: nearly everyone, in both arms and both countries, was in the second or third grade. Seven years later, most participants in intervention villages in both countries have progressed roughly a grade per year since then, while attainment in control villages has spread out, reflecting some combination of dropout, interrupted spells, and grade repetition. Figure 5 shows the resulting distributions. In Guinea-Bissau, the control distribution bunches at grades five and six – the last grades typically available in a participant's own village; continuing past them requires moving or traveling to another town. We suspect this is why the treatment effect is higher in Guinea-Bissau than in The Gambia; we note, however, that the highest grade reached by treatment village trial participants in both countries is roughly the same: 8.2 in The Gambia, and 7.8 in Guinea-Bissau.

The two countries differ in whether these gains are still accumulating. In The Gambia, participants from intervention villages are 14.2 percentage points more likely to be enrolled in school at the time of the survey (control mean: 65.8 percent), and 12.4 percentage points more likely to have been enrolled the previous year, so the gap in attainment there is likely to continue to grow. In Guinea-Bissau, self-reported enrollment is high in both arms (control mean: 88.4 percent) and we cannot reject that the arms are equal; there, the difference in attainment instead reflects faster progression through school. We treat the Guinea-Bissau enrollment figures with some caution: enrollment is self-reported, and near-universal reported enrollment can mask large differences in

actual attendance. The expenditure results are consistent with the attainment results: households of intervention participants in Guinea-Bissau spend 14,896 CFA (*Communauté Financière Africaine* Francs, roughly 27 USD) more per year on the participant's education, an increase of roughly 60 percent over the control mean of 24,452 CFA – consistent with more of these participants having reached secondary school, which must be sought outside the village. We find no such difference in The Gambia.

Patterns of time use are consistent with the enrollment results. In The Gambia, participants in intervention villages devote 10.5 percentage points more of their daytime to school and studying (control mean: 75.8 percent), offset by less time on domestic work and other activities. In Guinea-Bissau, where reported enrollment is high in both arms, we see no meaningful differences in time use.

Finally, Table 3 reports migration and labor market outcomes. In Guinea-Bissau, participants from intervention villages are 15.2 percentage points more likely to have moved away from their home village (control mean: 34.7 percent). This is not, for the most part, migration for work: instead, it helps explain the schooling results above, in which continuing past primary school in Guinea-Bissau requires leaving the village. In The Gambia, where most children can reach and return from upper primary/middle schools on foot in a day, we see no such difference.

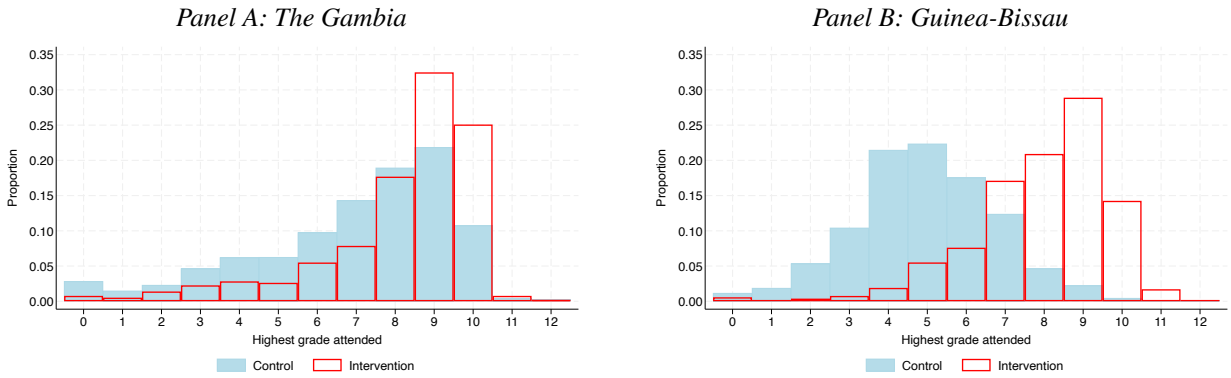
Turning to the labor market, we find little evidence of labor market participation for any trial participants. Under two percent of trial participants in The Gambia and under five percent in Guinea-Bissau have any paid work and, we find no evidence of differences in employment, hours, or earnings between treatment and control. Furthermore, the conditional estimates, shown in the bottom rows of Table 3, rest on a few dozen observations each. These participants are 16 to 19 years old and most are still in school; it is simply too early in their economic lives for the interventions' effects on labor market outcomes, if any, to appear. These outcomes are central to the longer-run agenda of this project, and we will return to them in future waves.

Table 3: Treatment effects on schooling, time use, migration, and labor market outcomes

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
Highest grade attended	1.194***	6.984	4,168	2.857***	4.969	1,649
	(0.132)			(0.203)		
Currently in school	0.142***	0.658	4,448	0.025	0.884	1,657
	(0.025)			(0.029)		
In school last year	0.124***	0.687	4,448	0.018	0.902	1,657
	(0.026)			(0.023)		
School expenditure	35	3,546	3,385	14,896***	24,452	1,390
	(146)			(3,163)		
Time: domestic work	-0.054***	0.137	2,846	-0.010	0.152	1,038
	(0.020)			(0.032)		
Time: paid work	-0.007*	0.013	2,846	-0.000	0.003	1,038
	(0.004)			(0.005)		
Time: school & studying	0.105***	0.758	2,846	-0.004	0.830	1,038
	(0.024)			(0.039)		
Time: play	-0.007**	0.013	2,846	0.022*	0.008	1,038
	(0.003)			(0.011)		
Time: other	-0.037***	0.079	2,846	-0.008**	0.007	1,038
	(0.010)			(0.003)		
Moved from home village	0.025	0.337	4,376	0.152***	0.347	1,689
	(0.026)			(0.051)		
Currently has paid work	-0.005	0.014	3,725	-0.002	0.044	1,652
	(0.003)			(0.016)		
Hours worked per week	-5.342	23.821	42	1.750	14.560	74
	(5.565)			(3.268)		
Earnings per week	-592	1,727	42	-3,011	8,225	74
	(424)			(3,547)		

Notes: Each row reports the results of a separate regression of the indicated outcome on an indicator for assignment to the intervention and fixed effects for the randomization strata, following Equation 1. School expenditure and earnings are reported in local currency units: Gambian dalasi in The Gambia (73 GMD $\approx$ 1 USD) and CFA francs in Guinea-Bissau (560 CFA $\approx$ 1 USD). Time use variables give the share of daytime activity devoted to each category. Hours worked and earnings are measured only among the small number of participants with paid work. *, **, and *** denote significance at the 10, 5, and 1 percent levels. Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum.

Figure 5: Distribution of highest grade attended at follow-up, by treatment arm



Notes: This figure plots the distribution of the highest grade attended at the time of the follow-up survey, separately for participants in intervention and control villages.

6 Aspirations and beliefs

The primary outcome is an index of the participant's future-oriented beliefs and aspirations, created by binarizing their responses to the relevant beliefs and aspirations questions in the survey, harmonizing them so that positive values connote higher aspirations, and summing them. Secondary outcomes include individual responses on these participant aspirations questions, an index of caregiver aspirations for the participant, and individual responses on caregiver aspirations questions.

6.1 Results

Figure 6 plots mean levels of the aspirations index and its three components – aspirations for education, work, and migration – by country and treatment arm; Table 4 presents the corresponding regression estimates.

The banner finding of this section is that participants in intervention villages have higher hopes for the future. The treatment effect on the aspirations index, the pre-specified primary outcome for this family, is 0.106 in The Gambia (control mean: 0.609) and 0.110 in Guinea-Bissau (control mean: 0.662). These estimates are remarkably similar across the two countries, despite the differences in intervention design and setting.

The individual components show that this difference runs through all three domains we measure. Participants in intervention villages are roughly 15 percentage points more likely to aspire to post-

secondary education (treatment effects of 14.9 and 15.4 percentage points in The Gambia and Guinea-Bissau, respectively) and roughly 13 percentage points more likely to aspire to a high-wage job in an urban area (12.9 and 13.6 percentage points). Differences in aspirations to live away from the home region are smaller – roughly 4 percentage points in both countries – in large part because these aspirations are already widespread in the control group, where means are 75.7 in The Gambia and 88.4 percent in Guinea-Bissau. Again, the estimates are strikingly consistent across the two countries. These differences are also larger than the corresponding differences in aspirations measured at the endline of the original studies, suggesting that, as with learning, the gap has grown rather than faded in the years since our last contact.

Neither intervention targeted aspirations, mentioned future schooling or employment, or otherwise addressed what participants should hope for from their lives; both ended when participants were nine to twelve years old. The differences we observe at ages 16 to 19 are therefore downstream of the early gains in skills and the subsequent divergence in schooling careers documented in the previous two sections. A natural interpretation, consistent with the broader literature linking aspirations to forward-looking economic behavior, is that participants who can read, and who have progressed farther in school, perceive higher returns to continued investment in their own futures – and form their hopes accordingly (Fruttero, Muller and Calvo-González, 2025; Bernard et al., 2026). We note one caveat: these questions measure aspirations rather than expectations, and we cannot distinguish hopes from forecasts in these data.

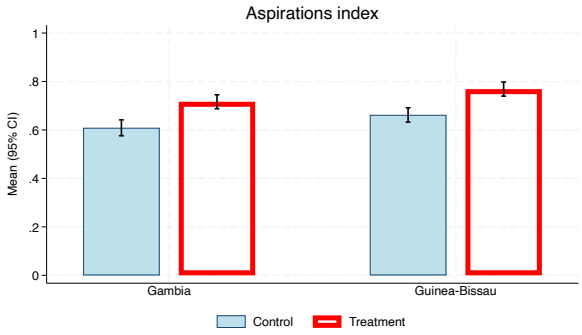
7 Family formation and gender norms

7.1 Outcome variables

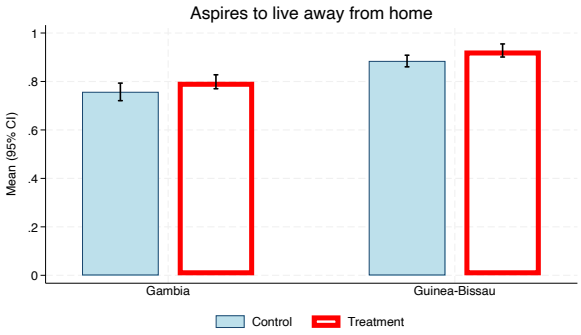
For this family of outcome variables, the primary outcome is an index of the participant's responses to three questions on gender norms, created by binarizing the responses, harmonizing them so that positive values connote more progressive viewpoints, and summing them. Secondary outcomes include individual level responses to each question and responses to a battery of questions about realized family formation and, if family formation has not begun, plans for family formation, which

Figure 6: Aspirations at follow-up, by treatment arm

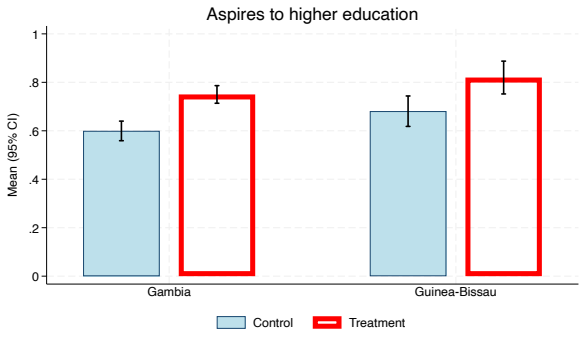
Panel A: Aspirations index



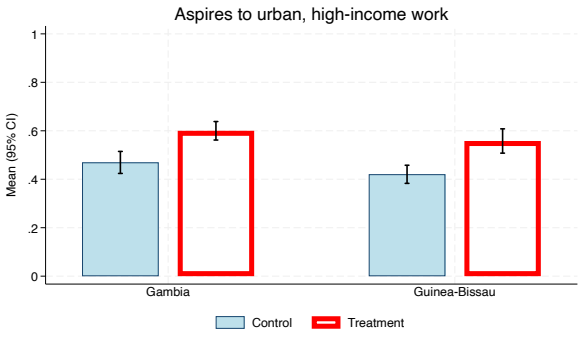
Panel B: Aspires to live away from home



Panel C: Aspires to higher education



Panel D: Aspires to urban, high-income work



Notes: This figure plots group means of the aspirations index (Panel A) and each of its binarized components (Panels B-D), by country and treatment arm, with 95 percent confidence intervals. Aspirations for education, migration, and work are coded as binary indicators for aspiring to post-secondary education, to living outside the home region, and to a high-wage job in an urban area, respectively; the index is the mean of the three.

Table 4: Treatment effects on aspirations

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
Aspirations index	0.106***	0.609	3,725	0.110***	0.662	1,652
	(0.021)			(0.020)		
Aspires to live away from home	0.041*	0.757	3,725	0.040**	0.884	1,652
	(0.022)			(0.018)		
Aspires to higher education	0.149***	0.600	3,725	0.154***	0.681	1,652
	(0.027)			(0.043)		
Aspires to urban, high-income work	0.129***	0.469	3,725	0.136***	0.420	1,652
	(0.029)			(0.032)		

Notes: Each row reports the results of a separate regression of the indicated outcome on an indicator for assignment to the intervention and fixed effects for the randomization strata, following Equation 1. Each aspiration variable is a binary indicator coded so that one connotes higher aspirations; the index is the mean of the three components. *, **, and *** denote significance at the 10, 5, and 1 percent levels. Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum.

are direct analogs to the realized family formation outcomes.

7.2 Results

Table 5 presents treatment effects on gender norms and on completed and intended family formation. The primary outcome, the gender norms index, is captured on a scale of 0 to 1, with 1 being most progressive and 0 being least. In it, we find a large difference in Guinea-Bissau but not in The Gambia. In Guinea-Bissau, the treatment effect is 0.087 (control mean: 0.575), and the action comes from the polygamy questions: participants in intervention villages are 17.5 percentage points less likely to accept polygamy for themselves (control mean: 59.3 percent do not accept) and 10.9 percentage points less likely to approve of polygamy for others (control mean: 51.0 percent). Views on whether sons should do housework are relatively progressive in both arms and both countries – control means of 70.7 and 62.1 percent – and do not move. In The Gambia, the estimate on the index is a precisely estimated zero (0.011, on a control mean of 0.300). The control means point to one reason for this divergence: in The Gambia, fewer than 15 percent of control participants reject polygamy for themselves, and fewer than 5 percent disapprove of it for others, leaving little margin along which to move. This is consistent with our observation, over more than a decade of

Table 5: Treatment effects on gender norms and family formation

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
			Gender norms			
Gender norms index	0.011	0.300	3,725	0.087**	0.575	1,652
	(0.011)			(0.034)		
Son should do housework	0.016	0.707	3,725	-0.024	0.621	1,652
	(0.035)			(0.032)		
Does not accept polygamy for self	0.018	0.148	3,725	0.175***	0.593	1,652
	(0.020)			(0.042)		
Does not approve of polygamy for others	-0.002	0.045	3,725	0.109*	0.510	1,652
	(0.009)			(0.061)		
			Marriage and fertility			
Currently or previously married	-0.008	0.023	3,725	0.004	0.013	1,652
	(0.006)			(0.007)		
Age at marriage	0.847**	15.896	75	-0.659	17.077	21
	(0.339)			(0.734)		
Spouse's schooling level (0–6)	0.545*	1.273	67	0.311	1.667	13
	(0.307)			(0.405)		
Has children	-0.005*	0.011	3,725	-0.001	0.023	1,652
	(0.003)			(0.008)		
Age at first child	-0.279	16.727	31	-0.962	19.577	39
	(0.448)			(1.556)		
Intends to marry	0.006*	0.988	3,651	0.002	0.984	1,629
	(0.003)			(0.006)		
Intended age at marriage	0.327	25.477	3,601	0.381	25.875	1,597
	(0.244)			(0.311)		
Wants children	0.002	0.993	3,694	-0.000	0.987	1,613
	(0.002)			(0.005)		
Desired age at first child	0.455*	27.335	3,654	0.041	27.074	1,577
	(0.235)			(0.333)		
Desired number of children	-0.272	6.869	3,280	-0.684***	4.902	1,571
	(0.203)			(0.240)		

Notes: Each row reports the results of a separate regression of the indicated outcome on an indicator for assignment to the intervention and fixed effects for the randomization strata, following Equation 1. The polygamy variables are coded as (1 − accepts polygamy for others/self), so that positive values connote more progressive viewpoints for all variables shown; the gender norms index is constructed from these binarized responses as described in the text. Age at marriage, spouse's schooling, and age at first child are measured only among the small number of participants who have married or had children. *, **, and *** denote significance at the 10, 5, and 1 percent levels. Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum.

work in the country, that the normative environment in The Gambia has, if anything, grown more religiously conservative over the study period – a rising tide of a different sort, working against any liberalizing effect of the intervention.

Turning to family formation, the most important fact in the completed-outcomes panel is how little family formation has yet occurred: roughly 2 percent of participants have ever been married and roughly 1 to 2 percent have children. As with labor market outcomes, these participants are 16 to 19 years old, and our data suggest that it is simply too early for us to observe effects on completed family formation. Because our estimates condition on marriage or childbearing – for example, the 0.85-year delay in age at marriage in The Gambia – they rest on samples of a few dozen participants in either country, and we interpret them with corresponding caution. Nonetheless, these outcomes are central to the longer-run agenda of this project, and we will return to them in future waves.

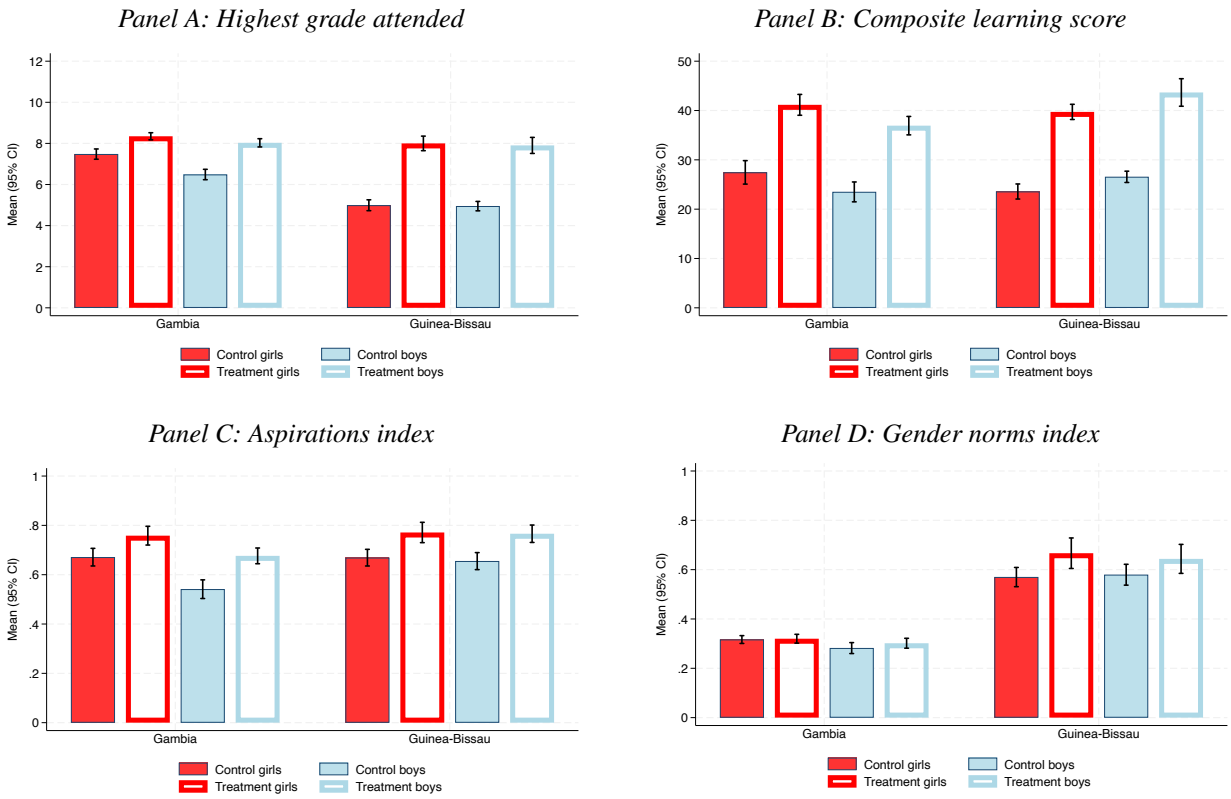
Intentions, which we can measure for the full sample, tell a more complete story. Marriage and childbearing are near-universal goals in both arms and both countries. Where we see differences, they point toward later and smaller families: in Guinea-Bissau, participants in intervention villages desire 0.68 fewer children (control mean: 4.90); in The Gambia, they desire their first child 0.46 years later (control mean: 27.3 years), though the latter estimate is only marginally significant.

8 Heterogeneity

We present analyses investigating the possibility of heterogeneity of treatment effects on the study's four primary outcomes. First, we present analysis by participant gender in Figure 7. This plots means of the four primary outcomes – highest grade attended, the composite learning score, the aspirations index, and the gender norms index – by country, treatment arm, and participant gender, along with relevant confidence intervals. Recent work has shown that, at least in education, general interventions often serve girls as well as or better than interventions targeted specifically at girls (Evans and Yuan, 2022*b*). Our results are consistent with that reading: for learning, gender norms, and aspirations, the treatment effects are all of comparable magnitude for girls and boys in both countries. For learning, girls in intervention villages score as high as, or higher than, boys in control villages. In The Gambia, boys seem to gain more from the intervention than girls in terms

of highest grade reached, but this serves more to close some of the existing gap between boys and girls, than it does to accelerate boys ahead of girls. For the gender norms index, the differences by participant gender are small and not statistically distinguishable.

Figure 7: Primary outcomes by participant gender, country, and treatment arm



Notes: This figure plots group means of each of the study's four primary outcomes, by country, treatment arm, and participant gender, with 95 percent confidence intervals.

We present our analyses for all pre-specified dimensions of heterogeneity in tabular form in Tables A6-A12. Aside from the gender results, we find little evidence of heterogeneity in impacts of the interventions. This, too, mirrors results from the original studies.

9 Conclusion

We estimate the long-term impacts of two randomized trials of a foundational literacy and numeracy program in The Gambia and Guinea-Bissau. The interventions combined structured pedagogy and teacher coaching, and were implemented at a high level of fidelity. In data collected seven years

after the intervention concluded, we document that the original learning gains in both literacy and numeracy not only persisted but grew over that seven-year period. These gains occurred despite the fact that no intervention was administered over the intervening time. We show that those foundational skills also aided in far greater levels of acquisition of many higher-level reading and math skills not taught in the original interventions. In addition, we see impacts on years of schooling and aspirations in both countries, and some evidence of an impact on gender norms in one of the two. Because the original intervention was randomized, we conclude that all of these differences are salutary effects of being in a village that was offered the original interventions, despite the fact that each focused exclusively on foundational literacy and numeracy.

Our study has a few key limitations worth noting. The first is that this was an evaluation of two very expensive interventions. While such interventions are increasingly within the set of policy options that many governments face, they nonetheless are extremely resource intensive and, as a result, may not be within the choice set of many interested governments. In ongoing research, we are evaluating whether or not the key insights from this bundle of interventions can be integrated into government school systems in such contexts, and if so, whether they will yield similar learning gains. The second limitation is an absence of administrative data to which to pair these findings. Because of the low capacity of both contexts and because a large proportion of individuals don't go on to take higher-level exams, we do not have school-level exams or other third-party data sources to complement the findings from our own primary data collection.

For policy, we believe there are a few key implications of our study. The first and the most important to us is that, in attempting to generate large learning gains in contexts such as these, where government capacity is very low, the size and impact of an intervention are likely proportional to its cost. Put more simply and directly, we think the key insight from our work, alongside several studies of interventions which try to bundle tools that are known to work in isolation to generate large learning gains, is the simple adage that "you get what you pay for." While the gold standard sought by many in this space is an intervention that yields such gains for $5 or $10 per child per year, all of the interventions which yield gains that are larger than a few tenths of a standard deviation

cost an order of magnitude more than that. While there are large public finance implications for this, we think that the problem of how to help children in these contexts reach adulthood with the ability to read with comprehension and do basic math is no less important than other global problems requiring international coordination, such as climate change and deforestation.

For future research, there are two obvious paths forward. The first is the one that we ourselves are pursuing: to estimate the longer-term impacts of such large learning gains on the economic and social lives of these individuals and their communities. We are particularly interested to see whether these gains, both in foundational literacy and numeracy and in higher-level skills, translate into greater labor market attachment and what kind of returns they yield in the labor market, given a large number of children who benefited from these. We also think there is potential for positive spillovers from having a critical mass of highly skilled people in contexts where previously this was outside the realm of possibility. The second path forward is to determine whether there are other configurations of this model of intervention which can yield similar gain for a smaller price. While we think it unlikely that such a gain can be reached for $5 or $10 per child per year, the hope of getting it for a somewhat larger sum, say $50 per child per year, is one that many parties are pursuing, and which holds great potential for improving welfare in these settings.

References

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

Asher, Sam, Paul Novosad, and Charlie Rafkin. 2024. "Intergenerational mobility in India: New measures and estimates across time and social groups." *American Economic Journal: Applied Economics*, 16(2): 66–98.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A multifaceted program causes lasting progress for the very poor: evidence from six countries." *Science*, 348(6236).

Barham, Tania, Karen Macours, and John A Maluccio. 2024. "Experimental evidence from a conditional cash transfer program: schooling, learning, fertility, and labor market outcomes after 10 years." *Journal of the European Economic Association*, 22(4): 1844–1883.

Barrera-Osorio, Felipe, Andreas de Barros, and Deon Filmer. 2024. "Long-term impacts of primary school scholarships: Evidence from Cambodia." *Journal of Policy Analysis and Management*, 43(1): 10–38.

Becker, Gary S. 1994. *Human Capital.* Chicago, IL:University of Chicago Press.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika*, 93(3): 491–507.

Bernard, Tanguy, Stefan Dercon, Kate Orkin, Giulio Schinaia, and Alemayehu Seyoum Taffesse. 2026. "The future in mind: Aspirations and long-term outcomes in rural Ethiopia." *The Quarterly Journal of Economics*, qjag002.

Birnbaum, Allan. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores*. , ed. Frederic M. Lord and Melvin R. Novick, 395–479. Reading, MA:Addison-Wesley.

Black, Sandra E, and Paul J Devereux. 2011. "Recent developments in intergenerational mobility." *Handbook of labor economics*, 4: 1487–1541.

Bock, R. Darrell, and Robert J. Mislevy. 1982. "Adaptive EAP Estimation of Ability in a Microcomputer Environment." *Applied Psychological Measurement*, 6(4): 431–444.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. "Enrollment without learning: teacher effort, knowledge, and skill in primary schools in Africa." *Journal of Economic Perspectives*, 31(4): 185–204.

Boone, Peter, Ila Fazzio, Kameshwari Jandhyala, Chitra Jayanty, Ganghadar Jayanty, Simon Johnson, Vimala Ramachandran, Ana Filipa Silva, and Zhaoguo Zhan. 2014. "The surprisingly dire situation of children's education in rural West Africa: results from the CREO study in Guinea-Bissau (Comprehensive Review of Education Outcomes)." In *African Successes, Volume II: Human Capital*. 255–280. Chicago, IL:University of Chicago Press.

Cascio, Elizabeth U, and Douglas O Staiger. 2012. "Knowledge, tests, and fadeout in educational interventions." *National Bureau of Economic Research Working Paper Number 18038*.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in action: teacher and health worker absence in developing countries." *Journal of Economic Perspectives*, 20(1): 91–116.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–2679.

Dubeck, Margaret M., and Amber Gove. 2015. "The early grade reading assessment (EGRA): its theoretical foundation, purpose, and limitations." *International Journal of Educational Development*, 40: 315–322.

Duflo, Esther, Pascaline Dupas, Elizabeth Spelke, and Mark Walsh. 2026. "Intergenerational Impacts of Secondary Education: Experimental Evidence from Ghana." *American Economic Review*, 116(9): 3586–3621.

Eble, Alex, and Maya Escueta. 2023. "When your bootstraps are not enough: How demand and supply interact to generate learning in settings of extreme poverty." National Bureau of Economic Research Working Paper 31388.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, et al. 2021. "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in The Gambia." *Journal of Development Economics*, 148: 102539.

Embretson, Susan E., and Steven P. Reise. 2000. *Item Response Theory for Psychologists.* Mahwah, NJ:Lawrence Erlbaum Associates.

Evans, David K., and Anna Popova. 2016. "What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews." *The World Bank Research Observer*, 31(2): 242–270.

Evans, David K., and Fei Yuan. 2022*a*. "How big are effect sizes in international education studies?" *Educational Evaluation and Policy Analysis*, 44(3): 532–540.

Evans, David K, and Fei Yuan. 2022*b*. "What we learn about girls' education from interventions that do not focus on girls." *The World Bank Economic Review*, 36(1): 244–267.

Fazzio, Ila, Alex Eble, Robin L Lumsdaine, Peter Boone, Baboucarr Bouy, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Simon Johnson, and Ana Filipa Silva. 2021. "Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau." *Journal of Public Economics*, 199: 104385.

Fruttero, Anna, Noël Muller, and Óscar Calvo-González. 2025. "The power and roots of aspirations." *The World Bank Research Observer*, 40(1): 58–103.

Glewwe, Paul. 2002. "Schools and skills in developing countries: education policies and socioeconomic outcomes." *Journal of Economic Literature*, 40(2): 436–482.

Glewwe, Paul, and Karthik Muralidharan. 2016. "Improving education outcomes in developing countries: evidence, knowledge gaps, and policy implications." In *Handbook of the Economics of Education*. Vol. 5, 653–743. Amsterdam, Holland:Elsevier.

Grant, Monica J, and Jere R Behrman. 2010. "Gender gaps in educational attainment in less developed countries." *Population and development review*, 36(1): 71–89.

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W Ozier. 2022. "Can education be standardized? Evidence from Kenya." *Becker Friedman Institute for Economics Working Paper*, , (2022-68).

Hanushek, Eric A, and Ludger Woessmann. 2008. "The role of cognitive skills in economic development." *Journal of Economic Literature*, 46(3): 607–668.

Hart, Emma R., Drew H. Bailey, Sha Luo, Pritha Sengupta, and Tyler W. Watts. 2024. "Fadeout and Persistence of Intervention Impacts on Social-Emotional and Cognitive Skills in Children and Adolescents: A Meta-Analytic Review of Randomized Controlled Trials." *Psychological Bulletin*, 150(11): 1207–1236.

Imbens, Guido W., and Charles F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica*, 72(6): 1845–1857.

Kerwin, Jason T, and Rebecca L Thornton. 2021. "Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures." *Review of Economics and Statistics*, 103(2): 251–264.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. "The effect of teacher coaching on instruction and achievement: a meta-analysis of the causal evidence." *Review of Educational Research*, 88(4): 547–588.

Kremer, Michael, and Alaka Holla. 2009. "Improving education in the developing world: what have we learned from randomized evaluations?" *Annual Review of Economics*, 1(1): 513–542.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The challenge of education and learning in the developing world." *Science*, 340(6130): 297–300.

Lakshminarayana, Rashmi, Alex Eble, Preetha Bhakta, Chris Frost, Peter Boone, Diana Elbourne, and Vera Mann. 2013. "The support to rural India's public education system (STRIPES) trial: a cluster randomised controlled trial of supplementary teaching, learning material and material support." *PLoS ONE*, 8(7): e65775.

Lee, David S. 2009. "Training, wages, and sample selection: estimating sharp bounds on treatment effects." *The Review of Economic Studies*, 76(3): 1071–1102.

McEwan, Patrick J, Erin Murphy-Graham, David Torres Irribarra, Claudia Aguilar, and Renán Rápalo. 2015. "Improving middle school quality in poor countries: Evidence from the Honduran Sistema de Aprendizaje Tutorial." *Educational Evaluation and Policy Analysis*, 37(1): 113–137.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro Ganimian. 2019. "Disrupting education? Experimental evidence on technology-led education in India." *American Economic Review*, 109(4): 1426–1460.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2013. "Contract Teachers: Experimental Evidence from India." National Bureau of Economic Research Working Paper Number 19440.

Petscher, Yaacov, Alison M Mitchell, and Barbara R Foorman. 2015. "Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer-administered measure of vocabulary." *Reading and writing*, 28(1): 31–56.

Piper, Benjamin, and Margaret M Dubeck. 2024. "Responding to the learning crisis: Structured pedagogy in sub-Saharan Africa." *International Journal of Educational Development*, 109: 103095.

Piper, Benjamin, Stephanie Simmons Zuilkowski, and Abel Mugenda. 2014. "Improving reading outcomes in Kenya: first-year effects of the PRIMR initiative." *International Journal of Educational Development*, 37: 11–21.

Piper, Benjamin, Yasmin Sitabkhan, Jessica Mejía, and Kellie Betts. 2018. "Effectiveness of Teachers' Guides in the Global South: Scripting, Learning Outcomes, and Classroom Utilization." RTI International Occasional Paper RTI Press Publication OP-0053-1805, Research Triangle Park, NC.

Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning*. Center for Global Development.

Protzko, John. 2015. "The Environment in Raising Early Intelligence: A Meta-Analysis of the Fadeout Effect." *Intelligence*, 53: 202–210.

Rossiter, Jack, David K Evans, Susannah Hares, and Catherine Henny. 2026. "The Illusion of Comparability Among Standardised Effect Sizes: Why Education Evaluations Should Report Raw Effects." Center for Global Development Working Paper Number 748.

Sprenger-Charolles, Liliane. 2008. "The Gambia: Early Grade Reading Assessment." World Bank Policy Report.

van der Linden, Wim J., and Ronald K. Hambleton, ed. 1997. *Handbook of Modern Item Response Theory.* New York:Springer.

Woessmann, Ludger. 2016. "The importance of school systems: evidence from international differences in student achievement." *Journal of Economic Perspectives*, 30(3): 3–32.

Appendix Tables

Table A1: Proportion reached at follow-up, by arm and data collection instrument

	Gambia			*Guinea-Bissau*		
	Control	Treatment	Difference	Control	Treatment	Difference
Reached: surveys	90.4	94.4	3.98***	76.8	81.0	4.29
			(1.38)			(3.60)
			[0.005]			[0.239]
Reached: tests	82.5	86.0	3.46**	70.6	80.9	10.29***
			(1.70)			(3.53)
			[0.044]			[0.005]
Baseline sample	2,458	2,060		1,463	649	
Clusters	55	56		33	16	

Notes: this table shows attrition rates, by arm (treatment/control) and data collection instrument (assessment/survey). Observations are considered as reached if the following primary endpoint is observed: learning in the assessment and highest grade attained in the survey. differences between the arms within each country by instrument cell are shown, with standard errors underneath and a p-value of equality below that.

Table A2: Lee Bounds for primary outcomes

	Gambia			*Guinea-Bissau*		
	TE	Lower	Upper	TE	Lower	Upper
Highest grade attended	1.194***	1.107	1.477	2.857***	2.843	3.231
Composite learning score	0.134***	0.118	0.148	0.157***	0.135	0.195
Aspirations index	0.106***	0.095	0.116	0.110***	0.094	0.139
Gender norms index	0.011	0.005	0.023	0.087**	0.060	0.116

Notes: this table shows Lee Bounds calculated for each primary outcome.

Table A3: Item Response Theory literacy and numeracy scores

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
IRT Score - Literacy B only	0.671***	-0.314	3,799	1.061***	-0.385	1,558
	(0.072)			(0.064)		
IRT Score - Literacy B and C	0.670***	-0.313	3,799	1.047***	-0.381	1,558
	(0.072)			(0.067)		
IRT Score - Numeracy B only	0.688***	-0.321	3,799	0.958***	-0.334	1,558
	(0.070)			(0.061)		
IRT Score - Numeracy B and C	0.690***	-0.322	3,799	0.961***	-0.335	1,558
	(0.069)			(0.061)		

Notes: Scores are EAP estimates from a 2PL IRT model, calibrated separately by country on students with a non-missing composite score, standardized to mean 0 and SD 1 within country. Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum. * <0.10, ** <0.05, *** <0.01.

Table A4: Zero scores on assessments

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
			Literacy			
Literacy, Part A	-0.078***	0.100	3,818	-0.038***	0.048	1,571
	(0.012)			(0.008)		
Literacy, Part B	-0.037***	0.056	3,825	-0.164***	0.212	1,571
	(0.009)			(0.019)		
Literacy, Part C	-0.168***	0.874	3,819	-0.140***	0.973	1,571
	(0.023)			(0.026)		
			Numeracy			
Numeracy, Part A	-0.001	0.002	3,818	-0.002	0.002	1,562
	(0.001)			(0.002)		
Numeracy, Part B	-0.094***	0.152	3,825	-0.014***	0.014	1,566
	(0.015)			(0.005)		
Numeracy, Part C	-0.113***	0.929	3,818	-0.217***	0.953	1,563
	(0.016)			(0.028)		

Notes: Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum. * <0.10, ** <0.05, *** <0.01.

Table A5: Effects of interventions on non-cognitive/life skills assessment

	Gambia			*Guinea-Bissau*		
	Treatment effect	Control mean	Number of observations	Treatment effect	Control mean	Number of observations
			Life skills			
Life skills index	0.032	0.000	3,843	0.142**	0.000	1,571
	(0.051)			(0.069)		
Life skills – problem solving and decision making	0.032	0.000	3,843	0.292***	0.000	1,571
	(0.058)			(0.079)		
Life skills – critical and creative thinking	0.021	0.000	3,843	0.049	0.000	1,571
	(0.056)			(0.082)		
Life skills – teamwork and collaboration	0.042	0.000	3,843	0.086	0.000	1,571
	(0.046)			(0.062)		

Notes: Standard errors, shown in parentheses, are clustered at the cluster level. Controls: Gambia – region, distance to road; Guinea-Bissau – composite randomization stratum. * <0.10, ** <0.05, *** <0.01.

Table A6: Heterogeneity in treatment effects on primary outcomes, by participant gender

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.135***	0.255	3,795	0.163***	0.253	1,558
	(0.014)			(0.012)		
Intervention × Female	0.001 [q = 0.604]			-0.010 [q = 1.000]		
	(0.015)			(0.012)		
Female	0.039***			-0.030***		
	(0.010)			(0.006)		
Panel B: highest grade attained						
Intervention	1.535***	6.985	4,165	2.837***	4.969	1,649
	(0.157)			(0.215)		
Intervention × Female	-0.674*** [q ≤ 0.001]			0.043 [q = 1.000]		
	(0.166)			(0.152)		
Female	0.999***			0.054		
	(0.129)			(0.099)		
Panel C: aspirations index						
Intervention	0.133***	0.608	3,724	0.114***	0.662	1,652
	(0.024)			(0.023)		
Intervention × Female	-0.046* [q = 0.111]			-0.009 [q = 1.000]		
	(0.025)			(0.030)		
Female	0.132***			0.014		
	(0.016)			(0.017)		
Panel D: gender norms index						
Intervention	0.020	0.300	3,724	0.071*	0.575	1,652
	(0.015)			(0.039)		
Intervention × Female	-0.017 [q = 0.232]			0.033 [q = 1.000]		
	(0.015)			(0.036)		
Female	0.034***			-0.010		
	(0.010)			(0.025)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("Female," equal to one for female participants and zero for male), and their interaction ("Treatment × Female"), estimated separately by country. Controls are indicators for region and distance to the main road in The Gambia and randomization-strata fixed effects in Guinea-Bissau; standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). "Control mean" is the control-group mean of the outcome and "Number of observations" the estimation sample; both are reported once per panel because they are common to the three coefficients. The composite learning score is the share of test questions answered correctly; the aspirations and gender-norms indices are scaled to $[0, 1]$. The value in square brackets below the interaction's standard error is the sharpened two-stage false discovery rate (FDR) q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): it is the smallest FDR level at which the treatment-by-heterogeneity interaction would be declared significant, correcting for multiple testing across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the (uncorrected) p-value of each coefficient.

Table A7: Heterogeneity in treatment effects on primary outcomes, by baseline household wealth

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.137***	0.254	3,745	0.174***	0.250	1,346
	(0.012)			(0.013)		
Intervention × More wealthy	-0.002 [q = 1.000]			-0.012 [q = 1.000]		
	(0.014)			(0.018)		
More wealthy	0.032***			0.007		
	(0.011)			(0.007)		
Panel B: highest grade attained						
Intervention	1.222***	7.006	4,095	2.830***	4.959	1,418
	(0.143)			(0.246)		
Intervention × More wealthy	-0.074 [q = 1.000]			0.188 [q = 1.000]		
	(0.138)			(0.192)		
More wealthy	0.239**			0.005		
	(0.119)			(0.129)		
Panel C: aspirations index						
Intervention	0.098***	0.608	3,671	0.110***	0.661	1,419
	(0.022)			(0.026)		
Intervention × More wealthy	0.045* [q = 0.427]			0.005 [q = 1.000]		
	(0.025)			(0.032)		
More wealthy	0.015			0.000		
	(0.020)			(0.022)		
Panel D: gender norms index						
Intervention	0.007	0.299	3,671	0.103**	0.562	1,419
	(0.012)			(0.044)		
Intervention × More wealthy	0.014 [q = 1.000]			-0.015 [q = 1.000]		
	(0.020)			(0.038)		
More wealthy	-0.001			0.002		
	(0.015)			(0.022)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("More wealthy," equal to one for households in the wealthier baseline category, classified from the materials the dwelling is built of, and zero otherwise), and their interaction, estimated separately by country. Controls are indicators for region and distance to the main road in The Gambia and randomization-strata fixed effects in Guinea-Bissau; standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). "Control mean" and "Number of observations" are reported once per panel because they are common to the three coefficients. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

Table A8: Heterogeneity in treatment effects on primary outcomes, by region

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.118***	0.255	3,797	0.164***	0.253	1,558
	(0.021)			(0.016)		
Intervention × NBR/Tombali	0.025 [q = 1.000]			-0.011 [q = 1.000]		
	(0.027)			(0.021)		
NBR/Tombali	-0.022			0.010		
	(0.022)			(0.011)		
Panel B: highest grade attained						
Intervention	1.070***	6.984	4,168	3.061***	4.969	1,649
	(0.188)			(0.280)		
Intervention × NBR/Tombali	0.196 [q = 1.000]			-0.366 [q = 1.000]		
	(0.258)			(0.371)		
NBR/Tombali	-0.297			-0.077		
	(0.212)			(0.234)		
Panel C: aspirations index						
Intervention	0.099***	0.609	3,725	0.116***	0.662	1,652
	(0.035)			(0.026)		
Intervention × NBR/Tombali	0.011 [q = 1.000]			-0.020 [q = 1.000]		
	(0.044)			(0.038)		
NBR/Tombali	-0.065*			0.055*		
	(0.036)			(0.032)		
Panel D: gender norms index						
Intervention	0.001	0.300	3,725	0.125***	0.575	1,652
	(0.014)			(0.043)		
Intervention × NBR/Tombali	0.015 [q = 1.000]			-0.072 [q = 1.000]		
	(0.021)			(0.060)		
NBR/Tombali	-0.001			0.007		
	(0.015)			(0.032)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("NBR/Tombali," equal to one for participants in North Bank Region in The Gambia or Tombali in Guinea-Bissau, and zero for Lower River Region or Quinara, respectively), and their interaction, estimated separately by country. Controls are an indicator for distance to the main road in The Gambia and randomization-strata fixed effects in Guinea-Bissau; because the region indicator is the heterogeneity dimension here, it enters through the main-effect and interaction terms rather than as a separate control. Standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). "Control mean" and "Number of observations" are reported once per panel. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

Table A9: Heterogeneity in treatment effects on primary outcomes, by distance to the main road

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.138***	0.255	3,797	0.155***	0.253	1,558
	(0.017)			(0.017)		
Intervention × Far from road	-0.010 [q = 1.000]			0.001 [q = 1.000]		
	(0.027)			(0.019)		
Far from road	-0.005			-0.015		
	(0.021)			(0.013)		
Panel B: highest grade attained						
Intervention	1.142***	6.984	4,168	2.738***	4.969	1,649
	(0.180)			(0.337)		
Intervention × Far from road	0.131 [q = 1.000]			0.183 [q = 1.000]		
	(0.257)			(0.405)		
Far from road	-0.212			-0.201		
	(0.199)			(0.299)		
Panel C: aspirations index						
Intervention	0.111***	0.609	3,725	0.093***	0.662	1,652
	(0.029)			(0.030)		
Intervention × Far from road	-0.012 [q = 1.000]			0.030 [q = 1.000]		
	(0.042)			(0.042)		
Far from road	0.008			-0.019		
	(0.031)			(0.031)		
Panel D: gender norms index						
Intervention	0.017	0.300	3,725	0.070	0.575	1,652
	(0.014)			(0.042)		
Intervention × Far from road	-0.016 [q = 1.000]			0.011 [q = 1.000]		
	(0.024)			(0.061)		
Far from road	-0.003			-0.058		
	(0.018)			(0.041)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("Far from road"), and their interaction, estimated separately by country. In The Gambia "Far from road" equals one for participants living above the median distance to the main road and zero otherwise; in Guinea-Bissau it equals one where there is no road and zero otherwise. Controls are an indicator for region in The Gambia and randomization-strata fixed effects in Guinea-Bissau; because distance to road is the heterogeneity dimension here, it enters through the main-effect and interaction terms rather than as a separate control. Standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). "Control mean" and "Number of observations" are reported once per panel. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

Table A10: Heterogeneity in treatment effects on primary outcomes, by caregiver education

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.139***	0.254	3,745	0.155***	0.253	1,555
	(0.013)			(0.010)		
Intervention × Caregiver educated	-0.016 [q = 0.511]			0.011 [q = 1.000]		
	(0.017)			(0.013)		
Caregiver educated	0.032**			0.015**		
	(0.013)			(0.007)		
Panel B: highest grade attained						
Intervention	1.279***	7.006	4,095	2.903***	4.970	1,646
	(0.139)			(0.223)		
Intervention × Caregiver educated	-0.344** [q = 0.073]			-0.149 [q = 1.000]		
	(0.142)			(0.226)		
Caregiver educated	0.419***			0.234		
	(0.113)			(0.151)		
Panel C: aspirations index						
Intervention	0.116***	0.608	3,671	0.118***	0.662	1,649
	(0.024)			(0.019)		
Intervention × Caregiver educated	-0.035 [q = 0.340]			-0.027 [q = 1.000]		
	(0.026)			(0.028)		
Caregiver educated	0.057***			0.039**		
	(0.017)			(0.015)		
Panel D: gender norms index						
Intervention	0.012	0.299	3,671	0.094***	0.575	1,649
	(0.012)			(0.034)		
Intervention × Caregiver educated	-0.005 [q = 0.613]			-0.020 [q = 1.000]		
	(0.015)			(0.037)		
Caregiver educated	0.022**			0.030		
	(0.011)			(0.024)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("Caregiver educated," equal to one if the participant's caregiver has ever attended formal schooling and zero otherwise), and their interaction, estimated separately by country. Controls are indicators for region and distance to the main road in The Gambia and randomization-strata fixed effects in Guinea-Bissau; standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). "Control mean" and "Number of observations" are reported once per panel. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

Table A11: Heterogeneity in treatment effects on primary outcomes, by presence of a school in the village at baseline

	Gambia			*Guinea-Bissau*		
	Estimate	Control mean	Number of observations	Estimate	Control mean	Number of observations
Panel A: composite learning score						
Intervention	0.146***	0.255	3,797	0.160***	0.253	1,558
	(0.019)			(0.010)		
Intervention × School in village	-0.016 [q = 0.906]			—		
	(0.022)					
School in village	0.012			-0.022		
	(0.017)			(0.017)		
Panel B: highest grade attained						
Intervention	0.968***	6.984	4,168	2.863***	4.969	1,649
	(0.200)			(0.205)		
Intervention × School in village	0.302 [q = 0.906]			—		
	(0.250)					
School in village	-0.249			-0.153		
	(0.212)			(0.177)		
Panel C: aspirations index						
Intervention	0.080**	0.609	3,725	0.112***	0.662	1,652
	(0.034)			(0.020)		
Intervention × School in village	0.038 [q = 0.906]			—		
	(0.039)					
School in village	0.021			-0.044		
	(0.027)			(0.046)		
Panel D: gender norms index						
Intervention	0.022	0.300	3,725	0.087**	0.575	1,652
	(0.016)			(0.034)		
Intervention × School in village	-0.017 [q = 0.906]			—		
	(0.018)					
School in village	-0.002			-0.008		
	(0.012)			(0.025)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("School in village," equal to one if at least one school was present in the participant's village at baseline and zero otherwise), and their interaction, estimated separately by country. Controls are indicators for region and distance to the main road in The Gambia and randomization-strata fixed effects in Guinea-Bissau; standard errors, in parentheses, are clustered at the unit of randomization (the cluster in The Gambia, the village in Guinea-Bissau). The interaction is not identified in Guinea-Bissau, where every treated participant's village had a school at baseline, so that the interaction term is collinear with the treatment indicator; the corresponding cells are marked "—" and the Guinea-Bissau "Treatment" and "School in village" coefficients are those from the additive model. "Control mean" and "Number of observations" are reported once per panel. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions estimated for that country in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

Table A12: Heterogeneity in treatment effects on primary outcomes, by baseline caregiver aspirations (The Gambia only)

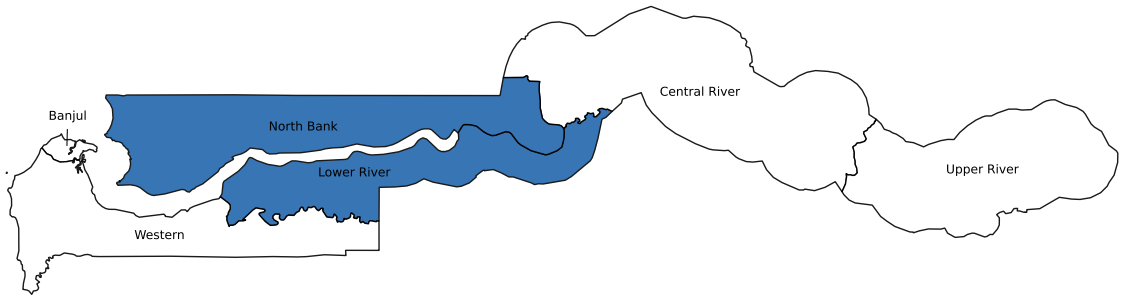
	Gambia		
	Estimate	Control mean	Number of observations
Panel A: composite learning score			
Intervention	0.135***	0.255	3,751
	(0.015)		
Intervention × High baseline aspirations	0.002 [q = 1.000]		
	(0.013)		
High baseline aspirations	0.052***		
	(0.009)		
Panel B: highest grade attained			
Intervention	1.219***	7.006	4,103
	(0.196)		
Intervention × High baseline aspirations	-0.003 [q = 1.000]		
	(0.193)		
High baseline aspirations	0.450***		
	(0.129)		
Panel C: aspirations index			
Intervention	0.120***	0.608	3,679
	(0.029)		
Intervention × High baseline aspirations	-0.018 [q = 1.000]		
	(0.024)		
High baseline aspirations	0.073***		
	(0.016)		
Panel D: gender norms index			
Intervention	0.020	0.299	3,679
	(0.014)		
Intervention × High baseline aspirations	-0.016 [q = 1.000]		
	(0.018)		
High baseline aspirations	0.017		
	(0.012)		

Notes: Each panel reports a single regression of the indicated primary outcome on an indicator for assignment to the intervention ("Treatment"), the heterogeneity indicator ("High baseline aspirations," equal to one if the caregiver's baseline aspiration for the participant's highest level of education was college or more, and zero if high school or less), and their interaction. Baseline caregiver aspirations were collected only in The Gambia, so the table has a single country panel. Controls are indicators for region and distance to the main road; standard errors, in parentheses, are clustered at the unit of randomization (the cluster). "Control mean" and "Number of observations" are reported once per panel. The value in square brackets below the interaction's standard error is the sharpened two-stage FDR q-value of Benjamini, Krieger and Yekutieli (2006), implemented as in Anderson (2008): the smallest FDR level at which the interaction would be declared significant, correcting across the four primary-outcome interactions in this table. * <0.10, ** <0.05, *** <0.01, based on the uncorrected p-value of each coefficient.

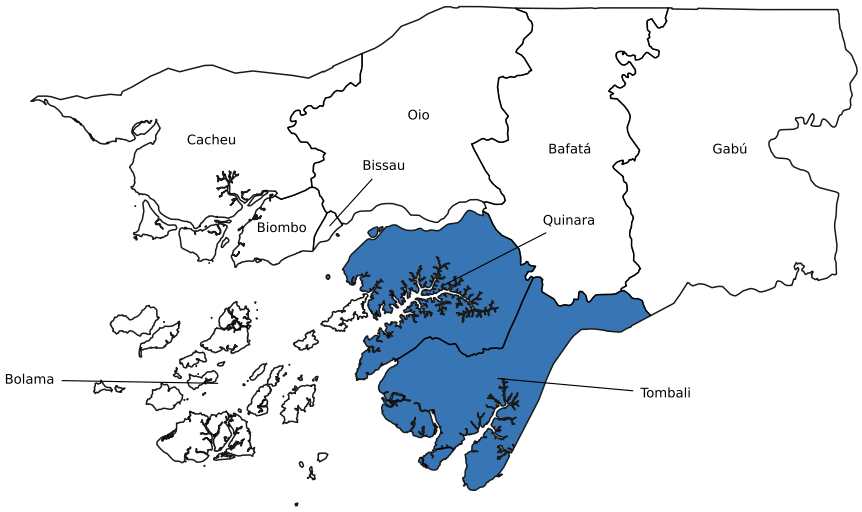
Appendix Figures

Figure A1: Location of project areas

(a) Gambia

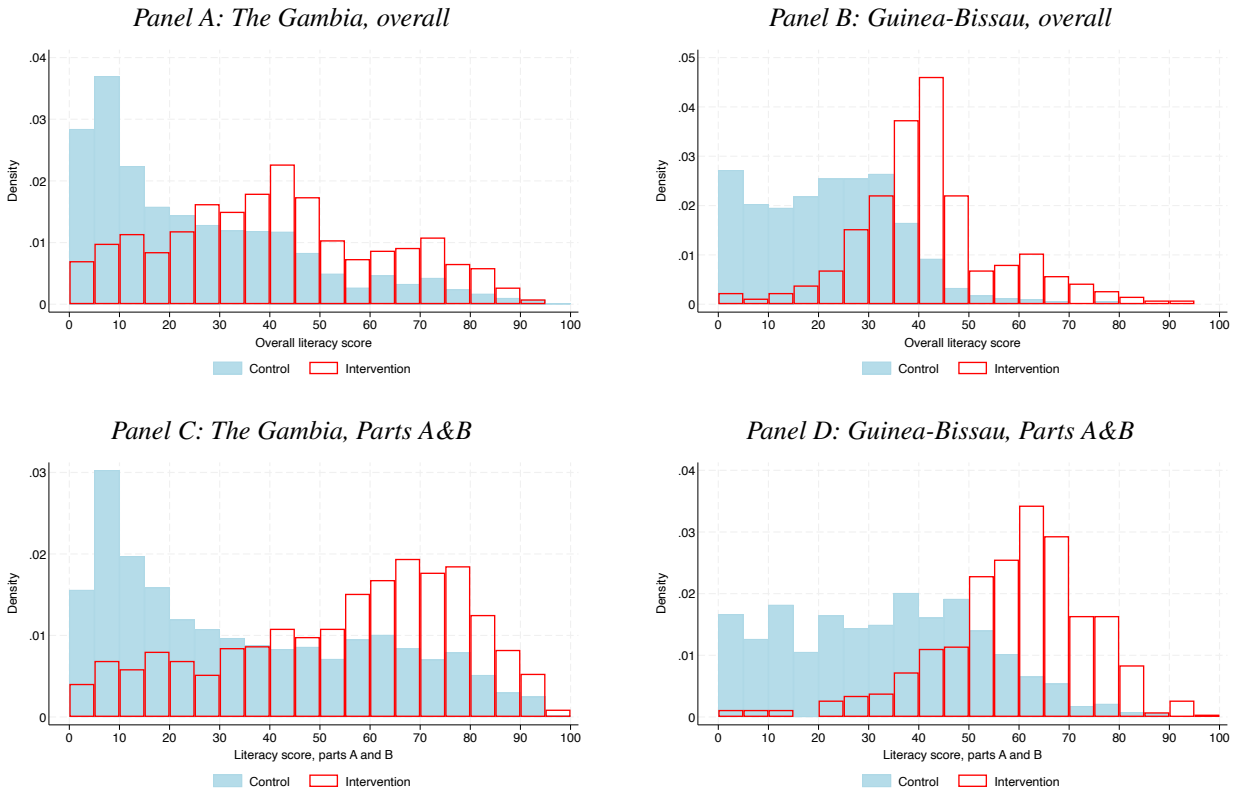


(b) Guinea-Bissau



Notes: This figure shows maps of the two countries in which our study takes place. In both countries, the map indicates the regions in which the work was done by shading that region blue.

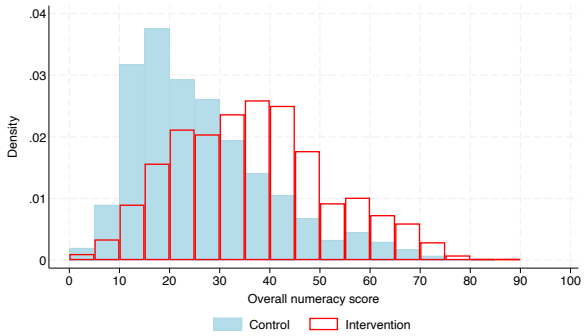
Figure A2: Composite literacy scores at follow-up, by treatment arm



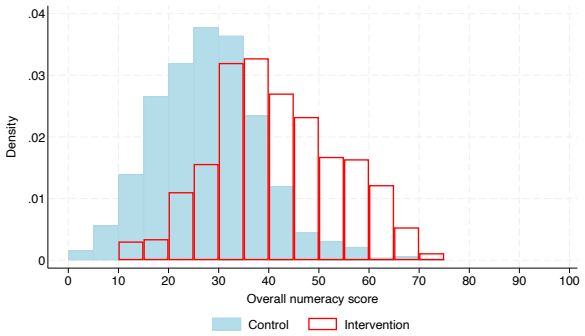
Notes: This figure plots the distribution of the literacy score at follow-up, separately for participants in intervention and control villages. Panels A and B show the overall literacy score (Parts A, B, and C); Panels C and D show the composite of Parts A and B only.

Figure A3: Composite numeracy scores at follow-up, by treatment arm

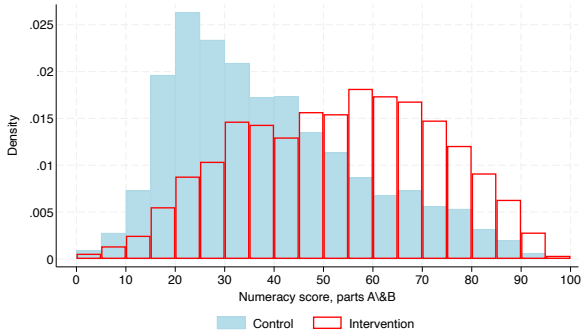
Panel A: The Gambia, overall



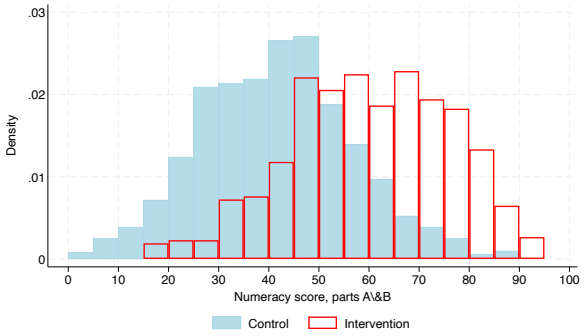
Panel B: Guinea-Bissau, overall



Panel C: The Gambia, Parts A&B



Panel D: Guinea-Bissau, Parts A&B



Notes: This figure plots the distribution of the numeracy score at follow-up, separately for participants in intervention and control villages. Panels A and B show the overall numeracy score (Parts A, B, and C); Panels C and D show the composite of Parts A and B only.

A Robustness: Construction of Item Response Theory scores

As a robustness check, we reconstruct our key learning outcomes using item response theory (IRT) in place of the percent-correct scores used in our main analysis. Rather than weighting every item equally, IRT estimates each child's latent ability from the full pattern of item responses, allowing items to differ in difficulty and in how sharply they discriminate between children of different ability (Embretson and Reise, 2000). We report these estimates because IRT scaling is a common expectation in the education-measurement literature; reassuringly, the resulting treatment effects, shown in Table A3, are very similar in magnitude and significance to our main estimates, indicating that our conclusions do not depend on the simpler scoring rule.

We estimate a two-parameter logistic (2PL) model (Birnbaum, 1968). Because every item is scored dichotomously, a logistic IRT model is the natural choice. We prefer the 2PL over the one-parameter (Rasch) model because the Rasch model constrains all items to be equally discriminating, an assumption that is implausible for assessments in which items vary widely in how well they separate stronger from weaker students; the 2PL relaxes this by estimating an item-specific discrimination parameter. We do not adopt the three-parameter (3PL) model: its additional guessing parameter is difficult to estimate stably, offers little benefit outside of large-scale multiple-choice testing, and would add noise rather than accuracy at our sample sizes. Person abilities are scored as expected a posteriori (EAP) estimates—the mean of each child's posterior ability distribution under a standard-normal population prior (Bock and Mislevy, 1982). EAP yields finite, well-behaved scores for every child, including those with perfect or zero raw scores, for which maximum-likelihood estimates are undefined; its only cost is a slight shrinkage of extreme scores toward the population mean, which is inconsequential for our purposes.

Because the two study contexts differ in language and instrument administration, we calibrate the item parameters separately within each country rather than imposing measurement invariance across them. Scores are estimated for children with a non-missing composite score and standardized to mean zero and standard deviation one within country, so that treatment effects are expressed in country-specific standard-deviation units, exactly as in our main specification. Items a child did

not attempt—including Part C, administered only to children who first cleared Part B—are treated as incorrect, consistent with our primary scoring and conservative with respect to the abilities of children who did not reach them.

We construct IRT scores only for Part B and for the combination of Parts B and C. We deliberately do not build a single IRT score spanning Part A as well: Part A measures distinct foundational skills, and folding clearly separate dimensions into one latent trait would violate the unidimensionality assumption on which IRT scaling rests (van der Linden and Hambleton, 1997). The close agreement between the Part B and the Part B-and-C estimates is consistent with these components tapping a common underlying skill.