Parents in the Classroom: Strengthening Government Capacity to Deliver Early Childhood Education*

Bet Caeyers[†] Lina Cardona-Sosa[‡] Sarah Cattan[§]

Sonya Krutikova[¶] Abu Siddique[‖]

Abstract

Preschool enrollment has expanded rapidly in low-income countries, but improvements in quality have lagged behind. This paper evaluates a scalable model that engages parents to improve early learning in contexts where both schools and homes are constrained. In collaboration with Ghana's Ministry of Education, we study the *Lively Minds* program in rural Ghana that trains mothers from the community to deliver structured, play-based learning activities in public preschools. Using a cluster randomized controlled trial, we find that the program increases children's cognitive development by 0.13 standard deviations and reduces problem behaviors, particularly among boys. The reductions in problem behaviors are driven by children of participating mothers, consistent with significant improvements in mothers' knowledge of child development and increases in developmentally supportive interactions at home. This pattern highlights the importance of engaging parents rather than relying exclusively on classroom-based inputs delivered by teachers or paraprofessionals. Notably, participation in the program has no adverse effects on mothers' well-being, and we show evidence that implementation quality and mothers' participation have been sustained as the program scaled to roughly one-third of rural districts nationwide in response to the findings of this evaluation. Overall, the results highlight parents as an effective and underutilized resource in poor rural settings for supporting early skill formation beyond the first 1,000 days and strengthening government capacity to deliver quality early childhood education.

JEL Codes: J13, I10, I20

Keywords: early childhood development, preschool quality, parenting, scalability.

*We are very grateful to *Lively Minds* NGO for their collaboration and support. We would like to acknowledge the contributions of Peter Leighton at an early stage of this project and of Salifu Amadu, Mubarik Ahmed, and the Innovations of Poverty Action team in Ghana during the data collection. We are grateful to Prof Elizabeth Spelke for sharing child assessment tools that she developed. This research was funded by the Global Innovations Fund, the Jacobs Foundation and the ESRC. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing up results. The study protocol was approved by the Ghana Health Service Ethics Review Committee (ref: GHSERC012/07/17) and IRBs at the Innovations for Poverty Action (ref: 14340) and the University College London (ref: 10167/001) in 2017. The trial was registered on ISRCTN (ISRCTN21215509) and AEA RCT Registry (AEARCTR-0002777) in March 2018.

[†]Chr. Michelsen Institute; IFS

[‡]World Bank

[§]IFS; IZA; CESifo; HCEO

[¶]University of Manchester; IFS

[‖]Royal Holloway University of London; IFS; IZA

1 Introduction

Early childhood investments have very high private and social returns, and early skill deficits are difficult and costly to remediate later in life (Heckman and Mosso, 2014; Currie and Almond, 2011; Heckman, 2012; Attanasio et al., 2022; Andrew et al., 2024). Yet governments in low- or lower-middle-income countries (LLMIC) often lack the resources and administrative capacity to deliver high-quality early childhood education at scale. Over the past two decades, enrollment in pre-primary education has increased substantially in LLMIC: from 13% in 2000 to over 30% in 2018 (UNICEF, 2019b). In many settings, however, such expansion has outpaced improvements in service quality: constrained budgets, weak implementation systems, and limited teacher capacity mean that children can spend years in school without mastering basic competencies (World Bank, 2018a). Therefore, scaling preschool access risks generating enrollment without learning, as classrooms become overcrowded and teachers remain poorly trained and under-resourced. At the same time, low levels of parental education and literacy mean that the main caregivers of young children, typically mothers, are often limited in capacity to provide cognitively rich inputs at home. Therefore, the main challenge for many LLMIC is how to improve early learning when both schools and homes are severely constrained.

We study this question in the context of Ghana's public preschool system. Ghana mandated two years of publicly provided pre-primary education in 2008 and now has one of the highest preschool enrollment rates in the region, serving close to 1.3 million children (Ministry of Education, Republic of Ghana, 2025). Despite near-universal enrollment, significant quality challenges persist, particularly in rural and high-poverty areas, including shortages of trained teachers, high teacher turnover and absenteeism, large class sizes, weak pedagogical practices, and limited funding (The World Bank, 2024). Many children also grow up in home environments that are not conducive to early learning, reflecting low parental education and time constraints, limited cognitive stimulation, and widespread use of harsh discipline (UNICEF, 2019a; Lim et al., 2023). In our baseline survey of 2,151 rural preschool-aged children, only 13% of caregivers reported any adult–child play in the previous three days. These constraints are reflected in low learning outcomes, with fewer than half of preschool-aged children developmentally on track in literacy and numeracy, and only about half of Primary 4 pupils meeting minimum competency benchmarks in English and mathematics (Lim et al., 2023; National Council for Curriculum and Assessment (NaCCA), 2023).

Against this background, we test a novel and scalable approach to improving children's learning and

development within this highly constrained setting being implemented by the Ghana Education Service (GES).[1] It is predicated on the idea that parents of young children are a key underutilized resource in the provision of quality early childhood care and education (ECCE) in LLMICs, particularly in rural communities, where needs are greatest. Accordingly, the Ghanaian Ministry of Education, working through the GES, is implementing a program developed by the NGO Lively Minds that trains and embeds mothers as regular volunteers within public preschools to deliver structured, play-based learning activities for preschool-age children.

We study the impact of this approach through a cluster-randomized controlled trial conducted across a sample of 80 preschools in two districts in of rural Northern Ghana, where the program was not yet operational. Half of the preschools were randomly selected for treatment, while the other half served as the control group and continued with standard preschool provision. In treatment preschools, the GES, with support from Lively Minds, recruited and trained mothers from the communities served by the preschools to deliver play-based learning activities during instructional hours. Interested mothers completed an eight-day training course led by preschool teachers, after which they were assigned to facilitate small-group play sessions (with activities such as counting, sorting, and puzzles) for up to two hours per week over 30 weeks of the school year. In addition to classroom involvement, participating mothers attended monthly parenting and well-being workshops run by the same preschool teachers.

Because the program is designed to change both what happens inside preschools and what participating mothers do at home, we frame its impacts on children's skills formation through two complementary channels: the preschool environment and the home environment. In the preschool, the program directly aims to improve the quality of the learning environment by introducing structured, play-based activities delivered in small groups by trained community mothers. This represents a direct investment in preschool-based early childhood education, which may improve cognitive skills through increased instructional time, more individualized attention, and exposure to play-based curricula (Grantham-McGregor et al., 1991; Currie, 2001; Hirsh-Pasek et al., 2009; Cascio and Schanzenbach, 2013). Structured, adult-supervised play and routines may also reduce problem behaviors among children by strengthening self-regulation, such as turn-taking and inhibitory control (Diamond and Lee, 2011; Healey and Healey, 2019). Beyond these direct classroom effects, a novel feature of this program is that it uses mothers as the marginal input in preschools, so implementation itself becomes an invest-

[1] The Ghana Education Service (GES) is the implementing agency of the Ministry of Education responsible for the delivery of public pre-primary, primary, and secondary education.

ment in parenting human capital with potential spillovers to the home environment: mothers acquire pedagogical knowledge and habits through training and repeated classroom practice, and can carry these skills back into the household, where they directly shape parenting inputs and investments in the child. This engagement may update maternal information and beliefs about returns to investment during early childhood (Cunha et al., 2022; Dizon-Ross, 2019). Therefore, if parental participation increases cognitively rich interactions and supportive parenting at home, children's skills should also improve through the home environment channel.

Our primary outcome of interest is children's cognitive and socioemotional development. Cognitive development is measured through direct one-to-one assessment by independent assessors and captures foundational literacy and numeracy skills, as well as executive functions. Socioemotional development is measured through direct assessment of emotional awareness as well as maternal report of problem behaviors. Our design also allows us to estimate the impacts of the program on the knowledge and practices of the participating mothers and the development of their children. To do so, identical mobilization and sign-up processes for participating in the program were conducted in the treatment and control communities. Comparing participating mothers and their children in the treatment preschools to those who signed up to participate in the following year but had not yet received the intervention in the control preschools allows us to isolate the causal effects of participation.

First, we find that early-childhood development (ECD) knowledge and skills of parents with minimal education and prior engagement in child development can be significantly improved by this relatively low-intensity program implemented within a highly constrained education system. In particular, we find that participating mothers' knowledge of child development increases by nearly half of a standard deviation (SD) and their knowledge of preschool quality by nearly a third of a SD; they also report substantially more engagement in play and learning activities with their child at home, and enumerator observations of mother-child play indicate a 0.26 SD increase in positive parenting behaviors. These effects are distinct from the broader effects of the program on parents in the community. While there is also some indication of improvement in ECD-related knowledge among non-participating mothers, this group primarily updates beliefs on easily observable aspects of preschool quality, not broader knowledge about child development. Non-participating mothers also do not increase the quality or quantity of engagement with their child and even show some evidence of reduced investments in play materials at home, perhaps perceiving these as less critical in light of improvements in quality at preschools.

Second, we find that following the training, engaging mothers in running small-group play-schemes during instructional hours in public preschools leads to economically meaningful gains in cognitive skills of the children enrolled in these preschools at baseline. Overall, there is a 0.13 SD increase in the aggregate cognitive development index, with improvements concentrated in emergent numeracy and executive function domains. The gains are broad-based across gender and are robust to alternative score construction that reduces measurement error. These average gains mask important distributional patterns: the program raises the probability that children score above the 25^{th}, 50^{th}, and 75^{th} percentiles of the control-group cognitive distribution, with little movement at the extreme tails.

The program also reduced behavioral difficulties among children. There is a marginally significant reduction of nearly 10% of a SD overall, which is within the range of average effects of 0.1-0.15 reported in meta-analysis of effects of preschool interventions on socioemotional skills (Holla et al., 2021). This effect is driven by a reduction specifically in externalizing behaviors (disruptive hyperactive behaviors) among those with a high problem load. It is concentrated among boys, who, in line with findings in the wider literature, are also more likely to exhibit externalizing behaviors. These impacts align well with the program's emphasis on structured, adult-supervised small-group activities and routines shown in the literature to be an effective approach to reducing externalizing behavioral difficulties.

Third, we find evidence that mobilizing parents as the program implementers generates additional benefits for the children of participating mothers, beyond the effects of improving classroom activities alone. Comparing children exposed to the play schemes whose mothers did not participate with children exposed to the same play schemes whose mothers did participate, we find that cognitive gains are similar across the two groups, but socioemotional improvements are concentrated among children whose mothers participated. We also find suggestive, though statistically imprecise, evidence of positive spillovers on the development of younger siblings of children whose mothers participate in the program.

The program is relatively low-cost at $37 per child (in implementation-year USD), as it relies on mothers volunteering their time and existing school infrastructure and staff rather than additional salaried staff or material inputs. This translates to roughly 1.6 equivalent years of schooling per $100 spent (Evans and Yuan, 2019). However, reliance on the existing workforce and volunteering mothers to deliver the program raises concerns about potential hidden costs. Participating mothers are predominantly illiterate women living in poor rural areas who tend to be disempowered and marginalized (see, for instance, Malapit and Quisumbing (2015)). Like many rural women in low-income settings, they face

substantial demands on their time from income-generating activities, farm work, and domestic responsibilities (Dinkelman and Ngai, 2022), making additional time demands potentially costly. Such reliance on low-status, poorly paid, overstretched female frontline workers and mothers is a common feature of ECD programs. In fact, there is a growing concern that the global push to expand early childhood provision may have unintended adverse consequences on the welfare of these already vulnerable groups (Ballard et al., 2023; Sullivan et al., 2024).

Our analysis offers encouraging evidence that this is not a concern with this program. We find no adverse effects of the program on the mental health and self-esteem of the participating mothers. This is consistent with the program's modest time demands on the participating mothers, only 1-2 hours per week, and emphasizing their well-being in the parenting workshops.

In sum, our findings show that a government-led model that mobilizes mothers as a low-cost input into public preschools can raise early human capital in settings where both school and home inputs are severely constrained. The program leads to gains in children's cognitive development, while further improvements in socioemotional outcomes are concentrated among children whose mothers participate in running the program, consistent with observed program-induced improvements in the ECD knowledge and practices of these mothers. These meaningful gains are achieved at relatively low cost and with no detectable reductions in the psychosocial well-being of participating mothers.

Related literature. We contribute to three strands of literature. First and foremost, we add to a nascent evidence base on scalable, cost-effective government-led approaches to improving early childhood education. While Ghana is a trailblazer, having been the first to provide universal access to public preschool in the region nearly 20 years ago, the challenges it faces with ensuring adequate quality are far from unique. We have seen these problems on a mass scale during the expansion of primary schooling (World Bank, 2018b) and are starting to see them at the preschool level as more countries join Ghana in expanding preschool provision (Blimpo et al., 2022; Bouguen et al., 2018). In all likelihood, as this continues, a growing number of countries will face the same challenges and require viable approaches to addressing them. The need for evidence on such approaches is all the more pressing as shrinking donor funding, rising debt, and growing student numbers are resulting in significant contractions of per-pupil education budgets across Sub-Saharan Africa, including Ghana (World Bank and UNESCO, 2024).

Existing evidence suggests that there are some promising approaches, but it is dominated by programs operated by nongovernmental organizations (e.g., see Dean and Jayachandran (2019); Martinez

et al. (2017)). Two very recent studies suggest that teaching assistant (TA) models can be an effective approach to improving the quality of government early childhood education provision, though accompanied by adequate training for the teachers to ensure that the TAs are deployed appropriately in the classroom (Ganimian et al., 2024; Andrew et al., 2024). In line with these studies, we also find that bringing additional support into the classroom can be an effective way of improving children's outcomes in government preschools. However, we provide evidence on the potential for parents, rather than additional salaried staff, to provide this support in a way that is not detrimental to the welfare of the participating parents. In terms of magnitudes, our 0.13 SD gain on an aggregate cognitive index is comparable to the ITT effect reported in recent TA intervention (Ganimian et al., 2024), which finds a 0.11 SD gain on a composite learning measure in household assessments. Our effects are also similar in magnitude to the estimated impacts from extensive-margin expansions in preschool access (Bassi et al., 2025, 0.16 SD gain on a skills index). The latter benchmark is informative because it suggests that, in low-capacity systems, relatively low-cost quality improvements within existing preschools (e.g., the *Lively Minds* program) can generate learning gains comparable to those achieved by large-scale expansions in preschool access.

Importantly, relative to this emerging evidence on what can work in government preschools, we also speak to another central question: whether such models can be sustained when shifted to routine delivery at scale. While our study offers results on a promising model, as is increasingly highlighted in the literature, scaling promising programs is fraught with challenges. There is a significant risk of "voltage drop" in effectiveness for a variety of reasons, including, among others, lack of suitably skilled workforce, weak incentives for maintaining quality in delivery, and the possibility that administrative and coordination constraints generate dis-economies rather than economies of scale (List et al., 2021). This risk is particularly salient in the delivery of ECD programs because of the critical importance that maintaining quality of delivery plays and evidence of loss of effectiveness of promising programs at scale (Andrew et al., 2018; Britto et al., 2018).

We contribute to this emerging scale-up literature by using rich administrative and monitoring data from the government roll-out of the *Lively Minds* program, which took place after the RCT that we study in this paper, allowing us to document mothers' participation, compliance, and delivery quality at scale. We are able to provide some encouraging, suggestive evidence that these risks can be mitigated in the implementation of this program.

This is because the Government of Ghana judged the early results from our study sufficiently promising to scale the program to 64 rural districts, reaching thousands of preschools and children by 2024. To mitigate key scaling risks, the government worked closely with Lively Minds NGO to strengthen quality assurance systems within the GES, reduce the NGO's role in day-to-day implementation, and increase non-financial incentives for teachers and participating mothers. Implementation data from 2,543 preschools across 46 districts between 2022 and 2024 indicate strong take-up and retention of participating mothers (on average 95% completed training and 86% remained active after two years), high compliance with core program components, and lower delivery costs at scale ($17.4 per child in 2025 USD, equivalent to just over $14 in 2018, which is less than half the RCT cost). Benchmarking these costs and impacts following the methodology in Angrist et al. (2025) suggests the program delivers just under one learning-adjusted year of schooling per $100 spent—placing it in the top half of the range of cost-effective education interventions documented in Angrist et al. (2025).[2] The program continues to expand, and key elements have been incorporated into national preschool reforms under the ongoing Ghana Accountability for Learning Outcomes Project (The World Bank, 2024).

Finally, by focusing on parents, our study contributes to the literature on parenting interventions. There is a large literature showing that parenting programs can be effective at improving parenting and child development for very young children (in the first 1,000 days) (Jeong et al., 2021; Carneiro et al., 2024). However, these programs tend to involve much more intensive training of the implementing workforce as well as parents themselves than the Lively-Minds model. Moreover, low-educated parents in settings such as ours may struggle more to support the development of early learning skills among preschool-age children than the development of children earlier on. Attempts to engage parents in the development of preschool-age children in poor, low-education settings have so far yielded mixed results. For example, (Angrist et al., 2025) evaluates an NGO-led program providing one-to-one training to parents of preschool-age children in a low-education setting but finds no impacts on parental knowledge or behavior. In comparison, (Özler et al., 2018) offers more encouraging evidence on a group-based parenting program for parents of preschool-age children. Our study adds to this evidence, demonstrating not only the feasibility of improving ECD knowledge and practices among parents of preschool-age children in poor, rural, low-education settings, but also doing so in a government-led way within the existing public education infrastructure.

[2]This is without accounting for additional socioemotional gains for children of participating mothers.

Structure. The paper is organized as follows: Section 2 describes the Ghanaian context and the L*ively Minds* program in detail. Section 3 outlines the experimental design, sampling, data, and empirical strategy. Section 4 presents the main results. Section 5 provides evidence on the program's opportunity costs. Section 6 assesses cost-effectiveness and documents the ongoing scale-up across 46 districts in Ghana. Section 7 concludes. All tables and figures are provided at the end.

2 Setting and intervention

2.1 Ghana Education System

Ghana introduced two years of publicly provided pre-primary education in 2008 and now has one of the highest preschool enrollment rates in West Africa, serving close to 1.3 million children nationwide (Ministry of Education, Republic of Ghana, 2025). Despite near-universal access, quality remains uneven, particularly in rural and high-poverty areas, where schools face persistent challenges including shortages of trained teachers, high turnover and absenteeism, large class sizes, and limited instructional resources (The World Bank, 2024). At the same time, many children grow up in home environments that provide limited cognitive stimulation, reflecting low parental education, time constraints, and widespread use of harsh discipline (UNICEF, 2019a; Lim et al., 2023). These constraints are reflected in weak early learning outcomes: fewer than half of preschool-aged children are developmentally on track in literacy and numeracy, and only about half of Primary 4 pupils meet minimum competency benchmarks in English and mathematics (Lim et al., 2023; National Council for Curriculum and Assessment (NaCCA), 2023).

The Government of Ghana has pursued a series of policy reforms aimed at strengthening early childhood education. For example, the Early Childhood Education Policy Framework published in 2021 provides a systematic, government-led structure for planning, curriculum implementation, teacher preparation, and quality assurance in KG services (**?**). Nevertheless, there remains a need for scalable, evidence-based interventions that can operationalize these policy frameworks. To date, much of the available evidence on effective ECD programs in contexts comparable to Ghana has come from small-scale initiatives implemented by non-governmental organizations outside of routine government systems.

Implementation of basic education in Ghana is administratively decentralized. While national policy and curriculum are set by the Ministry of Education and national agencies, day-to-day management and delivery of preschool services are largely the responsibility of district Ghana Education Service

(GES) offices. District directorates oversee teacher deployment, supervision, and school monitoring; organize in-service training; and are responsible for ensuring compliance with national guidelines. As a result, interventions to improve the quality of public preschools need to engage closely with the capacity, incentives, and implementation practices of district-level GES teams.

2.2 The *Lively Minds* Program

Against this backdrop, we evaluate the *Lively Minds* program, which is explicitly designed to strengthen the capacity of public preschools within Ghana's existing education system. Rather than operating as a parallel NGO-led model, the program is implemented through district GES structures, working with district teams, preschool teachers, and local communities to improve the quality of play-based learning in routine public preschool provision.

Program description. The *Lively Minds* program, developed by the Lively Minds NGO, seeks to improve the quality of public pre-school instruction by mobilizing and training women in the community to run weekly, play-based learning activities in classrooms alongside teachers. We refer to these women as "participating mothers" (PMs) since they tend to be mothers of preschool-age children. During each Play Scheme session, five mothers each facilitate a simple indoor "play station" (covering matching/sorting, numeracy, sizes/colours/senses, books, and building) using discovery-based activities in small groups of no more than five children. Preschool teachers supervise the sessions, while additional children participate in structured outdoor play led by another mother. The intended intensity is one indoor session per child per week, with outdoor play two to three times per week. In exchange for their time, participating mothers are offered monthly parenting and well-being workshops delivered by the preschool teachers.

Implementation model. The program is embedded within the Ghana Education Service (GES) and implemented at the district level. District GES teams are first oriented to the program, after which two teachers from each public preschool in the district receive a five-day training course on play-based pedagogy, classroom organization, and how to train participating mothers. Teachers then mobilize and train 30–40 mothers from the communities served by their preschools. This training consists of eight days of participatory workshops adapted for low-literacy participants. Following the training, PMs are divided into four groups, and each group is given a different day on which they teach at the KG for one hour (with extra time for preparation and debrief). During sessions run by the PMs in the preschools,

25 children are arranged in small groups (maximum of 5) and rotate around the following 5 indoor play stations: matching/sorting; numeracy; sizes, colors, senses; books; building. One PM runs each play station, and they teach using discovery-based teaching methods, rather than rote learning, which is the norm in formal education settings in Ghana. The remaining KG children play outdoor games, led by PMs. The KG teachers supervise the sessions.

In exchange for their time running the Play Schemes and to reinforce new behaviors, PMs are offered monthly parenting workshops by the preschool teachers following a curriculum developed by Lively Minds NGO on various topics related to child development and their own well-being.[3]

GES officials monitor the Play Schemes as part of their normal supervisory functions, and Lively Minds NGO conducts some additional unannounced monitoring visits. GES officials and Lively Minds NGO staff have monthly meetings to track the progress of the Play Schemes and identify corrective measures. There are monthly "top-up" training workshops for preschool Teachers where they discuss problems, share successes, and are trained to provide the PMs with a monthly Parenting Workshop.

Innovation. The *Lively Minds* model introduces a classroom experience that is markedly different from business-as-usual in rural Ghanaian pre-schools by combining small-group instruction, simple educational games, and classroom organization around rotating play stations. The program's design explicitly addresses common barriers to parental engagement: training is deliberately simple, highly practical, and demonstration-based (e.g., through role play), avoiding complex parenting curricula. A central aim is to build mothers' confidence in their ability to support children's learning despite limited formal education. The time burden is kept intentionally low—each participating mother volunteers for about one hour per week, rotating across four groups on different days; participation is encouraged through free monthly parenting and well-being workshops rather than financial payments. Finally, both the training and play curriculum were developed with and in rural communities of Northern Ghana and refined through nearly a decade of prior implementation, ensuring that they are context-appropriate and feasible in predominantly rural settings where women often have flexible farm-based work schedules.

[3]Topics include nutrition, hygiene, child rights, play, communication, malaria prevention, financial awareness, self-esteem, and inclusive education.

3 Study Design, Data, and Empirical Strategy

3.1 Experimental Design, Timeline and Sample

In order to evaluate the *Lively Minds* program, we conducted an RCT between 2017 and 2019. The randomization was conducted at the school level: 80 schools in 2 districts of Northern Ghana were randomized into two equally sized control and treatment groups. The selected districts had not yet introduced the *Lively Minds* program in their preschools. As described in the previous section, the business-as-usual implementation approach for this program is to implement it in all preschools once the district enrolls in the program. In order to ensure adequate power, implementation in the RCT districts was restricted to the schools randomly selected for treatment.

The RCT districts, Bongo and Tolon, are located in the Upper East and Northern regions, respectively. At the time of this study, these were among the most deprived areas of Ghana, with very high rates of poverty —50.4% in the Northern Region and 44.4% in the Upper East, compared to a national poverty rate of 16.5% — and some of the slowest long-run progress in poverty reduction since the early 1990s (Cooke et al., 2016). Consistent with these patterns, household survey evidence shows low levels of education, especially among women, in northern Ghana, with a high share of women reporting no formal schooling and low literacy (Ghana Statistical Service and UNICEF, 2018). Despite bordering each other, the two study regions differ along several dimensions, including language, religion, culture, and school quality, providing an opportunity to test the program in a heterogeneous set of communities that reflect the diversity of rural Ghana.

The randomization was conducted using two levels of stratification - circuit and school size.[4] The final sample includes 38 schools in the Bongo District, and 42 Schools in the Tolon District. Among the 40 schools allocated to the treatment group, 21 schools were located in Tolon and 19 schools in Bongo. The population of interest is children starting pre-school in the academic year 2017/18. In order to ensure that baseline data collection was completed before the start of intervention activities, sampling and baseline activities had to start before the beginning of the school year in July 2017. This meant that it was not possible to use school enrollment lists to identify the population of interest and do the sampling. We therefore conducted a census of households located in the vicinity of the sampled schools to identify those with children due to start pre-school. The households closest were enumerated first,

[4] A circuit is a geographical cluster of around 10 schools that falls under one supervisor from the Ghana Education Service (GES.

gradually moving to those further away to reach 150 households.[5] Of the 12,000 households listed in this way, a total of 4,486 eligible households were identified. We then drew a random sample of 2,407 children (approximately 30 per school) for the study.

Baseline data was collected over the course of three months between September and November 2017. The baseline survey included direct assessment of all of the children in the sample by trained interviewers, interviews with the primary caregiver of each of these children, the main KG teacher in each of the 80 KGs, as well as members of the community leadership team from the main community that each school served. Details of the baseline instruments can be found in (Amadu et al., 2018).

Endline data collection took place between September and November 2018, at the beginning of the next school year (2018/19) and following a two-month break in the Play Schemes and parenting workshops over the summer school holidays. Endline was completed before the program was rolled out to control schools in November 2018. We aimed to re-interview all of the children and primary caregivers in the baseline sample, including those who had moved to a different community since the baseline. We were able to collect complete endline data for 2,151 of the 2,407 children in our baseline sample, yielding an attrition rate of 11%, which is uncorrelated with treatment status (see Appendix Tables A1 and A2.)

Tables A3 and A4 present key baseline characteristics of non-attritors by treatment status. Table A3 does this for primary caregivers, while Table A4 for target children. The sample is well-balanced across key household, caregiver and child characteristics.

3.2 Ethics

We received ethical clearance for the project from the Ghana Health Service Ethics Review Committee (ref: GHSERC012/07/17), Innovations for Poverty Action (ref: 14340), and University College London IRB (ref: 10167/001).

3.3 Implementation and Compliance

All intervention activities took place between September 2017 and July 2018. Initial mobilization and introduction of district-level GES officials to the program took place in September 2017. The district team then implemented training of pre-school teachers in treatment schools. The teachers conducted

[5] See Amadu et al. (2018) for further details.

recruitment and training of participating mothers (PMs) in October 2017. A distinguishing feature of the Lively Minds program is its reliance on parents rather than teaching assistants, making it important to assess effects on both children and participating mothers. Because participation was voluntary, we constructed a comparable control group by replicating mobilization activities in control communities. In treatment areas, women were invited to participate in the current academic year, while in control areas, they were invited to sign up for the following year, when the program would be rolled out.[6]

Our main study sample includes 437 children whose mothers either ran the play-schemes in the treatment schools (253) or signed up to participate in the following year in the control group (184).[7]

The key assumption for identification of program impacts on PMs through comparison of the women who signed up to participate in the treatment and control schools is that selection into participation is the same in treatment and control preschools, even though in the latter sign-up was for a later start. We view this assumption as plausible for three reasons. First, the mobilization and sign-up activities were conducted by the same team in both treatment and control communities using an identical process, with the only difference being the timing of program rollout. Second, because children spend two years in pre-school and our sample consists of children entering their first year of KG, mothers in both treatment and control communities had a similar opportunity to participate while their own child was enrolled. Third, and most importantly, we verify that PMs in treatment and potential PMs in control are well balanced on baseline characteristics: Appendix Tables A5 and A6 show no systematic differences across a wide range of household, caregiver, and child characteristics within this PM sample. We therefore treat this sub-sample as a credible comparison group for estimating impacts on PMs.

Play Schemes started to run in treatment schools in November 2017 and continued during term time until the end of the school year in July 2018. The Play Schemes were designed to run 4 times per week in each school during school term with different children and mothers attending each session. Each child would participate in indoor games at least once per week and in outdoor play 2 or 3 times per week. Over the course of the academic year, the Play Schemes ran over three terms: first term 1st of November 2017 - 10th of December 2017 (6 weeks), second term 15th of January 2018 - 15th of April 2018 (13

[6]Our sample consists of children entering their first year of kindergarten (KG), which lasts two years in Ghana, so mothers in control communities still had an incentive to sign up even with a one-year delay.

[7]A total of 1,480 women signed up to participate in running the Lively Minds Play Schemes in the 40 treatment schools. The great majority of these women were mothers of KG-age children. Some were mothers of the children we sampled for the evaluation. To identify them, we matched the names of all primary caregivers in our sample to GES's lists of women participating in the Lively Minds program. A total of 253 mothers of children in our sample in the treatment group were matched. We then compared lists of mothers who signed up to participate in the following year in the control communities; 184 matched mothers of children in our sample were in the control group.

weeks) and, the last 14th of May 2018 - 24th of July 2018 (11 weeks). At full implementation, the Play Schemes would have operated for 30 weeks in total, providing each child with approximately 30 hours of participation over the academic year.

Implementation data collected by supervisors from the GES and Lively Minds NGO teams during random visits to preschools in the treatment group suggest that compliance was high. The Play Schemes were found to be running in 80% of all monitoring visits. The main reasons for the Play Scheme not running at a particular visit were bad weather, community funerals, other school/community event or low attendance of PMs, the latter being the most frequent reason recorded by the monitoring team. Using this as a proxy of the actual proportion of Play Schemes that were running suggests that children were exposed to an average of 23-25 hours of Play Schemes during the school year. On most monitoring visits, there were 2 teachers present, an average of 8 PMs, and 24 children. In most cases mats were organized as expected, and in more than half of the cases, the practice of discovery-based teaching was observed. In addition, each school received an average of between 4 and 5 monitoring visits at the monthly parent workshops organised by the preschool teachers as part of the program over the course of the academic year. In 88% of the visits, the workshop was taking place. In most cases, both teachers were present with an average attendance of 30 mothers per session.

Teacher data collected as part of the endline survey provides additional evidence on compliance. Table 1 shows that teachers in treatment preschools were almost five times more likely than control teachers to agree with the statement that parents actively contribute to preschool activities (Column 1), 38 percentage points more likely to report having received active parental support during the academic year (Column 2), and 29 percentage points more likely to report that toys, games, or other equipment were accessible to children during free play on a typical preschool day (Column 3).

3.4 Outcomes

We pre-specified child development in the cognitive and socioemotional domains as the primary outcome (see AEA RCT registry no. AEARCTR-0002777).

3.4.1 Child development

At both baseline and endline we administered the International Development and Early Learning Assessment (IDELA) to measure cognitive and socio-emotional development, alongside selected items from

the Strengths and Difficulties Questionnaire (SDQ) to capture behavioral difficulties (Goodman, 1997). The IDELA was administered directly to child by trained interviewers and covers emergent literacy, emergent numeracy, executive functioning (EF), and emotional awareness. The SDQ was administered to the child's primary caregiver to elicit reports of behavioural difficulties (Hoosen et al., 2018). At endline, we supplemented these measures with a set of child tasks developed at the psychology Laboratory of Development Studies at Harvard and piloted in Ghana (Coffey and Spelke, 2024). These tasks were added to increase the discriminatory power of our measures while mapping onto the same core domains captured by the IDELA and, like the IDELA, were administered directly to children by trained enumerators. The scores used in the main analysis are constructed following developer-prescribed scoring rules. Baseline scores of these measures are standardized to have mean=0 and standard deviation=1. Endline scores are standardized relative to the control group distribution. Appendix B provides further details about the assessments, construction of scores for the main analysis and robustness to alternative ways of scoring.

To study potential intra-household spillovers, we pre-specified developmental outcomes of the target child's siblings as secondary outcomes and collected child development data for one older and one younger sibling of each target child, where applicable. Younger siblings' development was captured using the Caregiver-Reported Early Development Index (CREDI) short form, a caregiver-report instrument designed for population-level measurement of early development from birth to age three in low-resource settings; scores were constructed following the developers' age-standardization guidance and then standardized relative to the control group (McCoy et al., 2017). Older siblings completed direct, one-to-one assessments consisting of a curriculum-aligned literacy and numeracy test developed for the Ghanaian context within the Ministry of Education, complemented by selected tasks from the battery developed by Harvard Laboratory of Development Studies and used with the target child. See Amadu et al. (2020) for further details.

3.4.2 ECD Knowledge, Practices and the Home Environment

The aim of *Lively Minds* is to improve child development through improving ECD care and education skills of parents and engaging them more actively in supporting the development of young children in their communities. Our pre-specified set of secondary outcomes, therefore, includes a rich set of measures capturing different dimensions of ECD knowledge and practices among parents of children in

the sample collected through interviews of the primary caregivers of the children in our sample. First, knowledge about ECD is assessed using a short subset of the Knowledge of Infant Development Inventory (KIDI), a widely used instrument that elicits caregivers' beliefs about children's developmental capabilities and appropriate expectations (MacPhee, 1981). Second, we measure caregivers' understanding of pre-school quality using a contextually tailored tool in which parents were shown illustrated vignette pairs depicting contrasting classroom practices and asked to select the option more conducive to child development.

We capture ECD practices by combining caregiver reports with direct observation. Our first measure of ECD practices was collected through observing how mothers interacted with their own children during structured play tasks and coding a range of supportive and less supportive instructional behaviors, drawing on items from the UNICEF Multiple Indicator Cluster Surveys (MICS) (Bornstein et al., 2010). We also capture ECD behavior through measuring material and time inputs into the child using the Family Care Indicators (FCI), which records the availability of play and learning materials in the home and recent play and learning activities with the child (Hamadani et al., 2010).

Because there is no developer-prescribed scoring rule for the full set of these measures, we construct our main indices of parental knowledge and behavior using Item Response Theory (IRT), while also reporting raw-score results in the appendix (as per the pre-analysis plan). See Appendix B for further information about these measures, how they were scored, and robustness to alternative scoring approaches.

3.5 Empirical strategy

The experimental design allows us to study the causal impact of *being eligible* for the *Lively Minds* program (intent-to-treat effect – ITT). In other words, we compare outcomes of children who were reported during the pre-baseline census as planning to attend one of the "treatment" pre-schools to those reported as planning to attend one of the "control" pre-schools. In line with the pre-analysis plan, we condition the outcomes at endline on a set of observable variables (including the outcome at baseline) in order to improve the precision of our estimates. Formally, we estimate the ITT effect of the *Lively Minds* program on the outcomes of interest by estimating the following regression:

$$Y_{is,1} = \alpha + \beta Treat_{s,0} + \sigma Y_{is,0} + \gamma X_{is,0} + \theta Strata_s + \delta District_s + \epsilon_{is,1} \tag{3.1}$$

where $Y_{is,1}$ is the outcome of interest for child i intending to go to pre-school s measured at endline; $Treat_{s,0}$ is a dummy equal to 1 if the pre-school s received the *Lively Minds* program; $Y_{is,0}$ is the same outcome measured at baseline[8]; as pre-specified $X_{is,0}$ is a set of child, household, community and school characteristics measured at baseline that were imbalanced across the treatment and control group samples at endline; $Strata_s$ is a fixed effect for the randomization strata of school s; $District_s$ is an indicator for whether the school is located in Bongo or Tolon district; finally $\epsilon_{is,1}$ is the random error term, clustered at the school level (our unit of randomization).

We estimate equation (3.1) by OLS so that β is the estimated average ITT impact of *Lively Minds* on outcome $Y_{is,1}$. Given the number of outcomes examined across cognitive subdomains, socioemotional subdomains, and parental inputs, we account for multiple inference using Romano-Wolf step-down adjusted p-values within pre-specified outcome families (Romano and Wolf, 2005). We report the unadjusted p-values in all of the tables and the adjusted p-values wherever they are relevant.

4 Results

4.1 Knowledge and practices of parents

The approach developed by *Lively Minds* is that it mobilizes low-educated, often illiterate, mothers as classroom facilitators. The key assumption underlying this model is that it is possible to train the mothers for this role through a short (8-day) group-based training course led by preschool teachers, supplemented by a monthly parenting workshop.

We, therefore, start by showing estimates of impacts of the *Lively Minds* program on (i) mothers' ECD knowledge, (ii) investments in child development at home, and (iii) quality of interaction with children. See Section 3 for details on how these are measured. As discussed in Section 3, we use outcome measures constructed using IRT in our estimates and present estimates using raw scores in Appendix A. We present impacts separately for mothers who directly participate in the program and those who do not, but whose children were registered as planning to go to a participating preschool during the baseline. We present impacts on this latter group to capture any indirect impacts of the program on mothers in the community. Differences in impacts and p-values for tests of statistical significance on these differences are shown at the bottom of the table.

[8]There are some outcomes for which there is no exact corresponding baseline measure. In these cases we add controls that are likely to constitute good proxies for a direct baseline measure of the outcome.

ECD Knowledge. Table 2, Panel A shows clear improvements in knowledge about child development and preschool quality among PMs as a result of the program. There are large and statistically significant treatment effects on both of these domains: knowledge about child development increases by just under half of a SD (column 1, $p = 0.016$), and knowledge about what makes a good pre-school by close to a third (column 2, $p = 0.062$).[9]

As discussed in Section 2, the approach developed by *Lively Minds* differs markedly from business-as-usual methods in Ghanaian public preschool by including small-group play, an emphasis on simple educational games with props, and classroom organization designed to support children's learning. It is thus plausible that even parents who are not directly participating in the program update their beliefs about child development and preschool quality. We see some evidence of this in Panel B of Table 2. While, in comparison to the PMs, we observe no effect on knowledge of non-PMs about child development (column 1, $p = 0.884$), there is a statistically significant increase of 0.17 SD (column 2, $p = 0.026$) in knowledge about preschool quality among non-PMs. This effect size for non-PMs is around half of that for PMs, though the difference in the effects is not statistically significant (column 2, Panel C, $p = 0.359$). These patterns are consistent with differences in exposure to program content by PM status. Non-PMs update beliefs that map to more easily observable changes in what happens at preschools as a result of the program, while PMs additionally benefit from training on child development.[10]

ECD practices. Better knowledge about ECD is important because it can lead to changes in parents' behavior that are beneficial for child development (Cunha et al., 2022; Carneiro et al., 2024). We now examine the impacts of *Lively Minds* on these behaviors. While the training, parenting workshops, and experience running the Play Schemes may have influenced the ECD practices of participating mothers, their low levels of education and the relatively light-touch nature of the intervention make impacts less certain.

Results in columns 5-6 of Panel A in Table 2 point to a significant improvement in the quality of PMs' engagement with children. This was measured through enumerators' observations of the quality of PMs' interactions with their own children (see Section 3 for details on this measure). We see a large

[9]Note that these estimates use IRT-constructed outcomes; the corresponding specifications using mean-of-raw score indices show very similar effects (Table A7).

[10]In addition to an overall improvement in their knowledge of child development, which is not there for the non-PMs, we also see improvements in knowledge about preschool quality on more dimensions among PMs than non-PMs (Table A8): non-PMs update primarily on knowledge about play-based learning (perhaps the most easily observed feature of *Lively Minds*), while PMs show broader updating, with improvements in knowledge about both play-based learning and what good school building should be like.

and significant increase of 0.26 SD in positive behaviors (supportive, responsive, and developmentally appropriate), which goes from being highly statistically significant to marginally significant once we adjust for multiple hypothesis testing (column 5, Panel A, $p = 0.116$). As shown in Appendix Table B22, once we relax the assumption of full measurement invariance in the construction of the IRT scores for this parenting measure, the effect size increases to 0.33 ($p = 0.007$).

Results in columns 3-4 of Panel A in Table 2 show further evidence of improvements in ECD practices of PMs. There is a significant increase in the frequency of play and engagement in other developmentally supportive activities with children of PMs by adults in the households of the PMs. Among the potential PMs in the control group, only 36% reported that anyone in the households had conducted any play activities with the sampled child in the last 3 days. Among PMs in the treatment group, this proportion increases by 50% (18 p.p.) (column 4, Panel A, $p = 0.012$). Effects are concentrated on the time rather than the material investment margin: there is no evidence of any change in investment in play materials for the children (column 3, Panel A, $p = 0.766$).

The large and significant increase in play activities without an accompanying increase in material investments aligns well with the adoption of low-cost, routine-based ways of engaging children that were emphasized through the training and parenting workshops; PMs may have implemented what they learned on embedding learning opportunities into everyday household chores. Disaggregating the results by specific play activities in Table A9, we see increases in reading to the child, telling them stories, playing, and drawing with them. We do not see impacts on singing, going out, or playing with objects.

In light of evidence of changes in PM parenting behavior as well as impacts on non-PMs beliefs about pre-school quality, it is possible that the *Lively Minds* program also had indirect impacts on the parenting of non-PMs. Column 3-6 in Panel B of Table 2 shows some suggestive evidence of this, though in a way that is very different from the PMs. Specifically, there is no impact on the time spent with children of non-PMs or the quality of engagement with children, in contrast to the large positive effects among PMs; the differences between treatment effects on these outcomes for PMs and non-PMs are statistically significant (columns 4-5, Panel C). There is, however, some evidence of a *reduction* in investment in play materials at home (column 3), which we do not see for the PMs. Among non-PMs, investment in play materials at home decreased by close to a fifth of a standard deviation ($p = 0.088$). Furthermore, this reduction is driven by negative impacts of the program on several items, including household objects used as toys, books, and toys for learning shapes (Table A9). The combination of

the effects on non-PMs suggests that they were aware of improvements that happened in the preschools because of Lively Minds and responded by reducing investments at home. It is plausible that non-PM households viewed the preschool program as a substitute for home inputs, reducing investments in play materials because they perceived that children's developmental needs were being met at school.

4.2 Child development

Cognitive skills. In line with evidence of the success of *Lively Minds* in training mothers to strengthen their ECD knowledge and practices, along with high compliance in the delivery of Play Schemes in the classrooms by the trained mothers, reported in Section 3, we find that the program produced economically meaningful gains in children's cognitive development. Table 3 shows an ITT treatment effect of 0.13 standard deviations (SD; column 1, $p = 0.022$) on the aggregate measure of cognitive development, which combines emergent numeracy, emergent literacy, and executive functions. Disaggregating by subdomain, the gains are concentrated in emergent numeracy (column 3, $p = 0.040$) and executive functions (column 4, $p = 0.034$).

The first row of Table 4 further shows that the program was not equally effective at all points in the cognitive skill distribution. It significantly increased the probability of getting a score above the 25^{th}, 50^{th}, and 75^{th} percentile of the control group distribution, with effects getting bigger at higher levels of skills, but had no impacts at the tails—on the probability of getting a score above the 10^{th} or the 90^{th} percentile. There is some indication that the distributional pattern of treatment effects is different from this for the executive functioning domain: intervention was most effective at increasing executive functions among those at the lower end of the executive functions distribution.

The program, however, was equally effective at improving the cognitive skills of boys and girls (column 1, Table 5). Estimated treatment effects are similar in magnitude for boys ('boy' row, $p = 0.036$) and girls ('girl' row, $p = 0.028$), and we cannot reject equality of effects across genders ('difference' row, $p = 0.767$). It was also equally effective for children of PMs and non-PMs (Table 6). This result suggests that the benefits of the program on children's cognitive skills are driven by changes that take place in the pre-schools as the result of the program - likely the introduction of the Play-Schemes - rather than any changes in the home learning environment of the children of PMs.

Socioemotional development. Table 7 reports effects of the *Lively Minds* program on children's socioemotional development. A positive coefficient in column 1 would indicate an improvement in emo-

tional awareness, and negative coefficients in columns 2–4 would imply a reduction in behavioral difficulties.

We find little evidence of program impact on emotional awareness (column 1, $p = 0.560$) and across the emotional awareness score distribution (see Table 4). This finding holds equally for boys and girls (see Table 5, column 2). There is, however, some evidence that the program reduces children's behavioral difficulties. Column 2 in Table 7 shows an effect of -0.098 SD, although the estimate falls just short of conventional significance levels (column 2, $p = 0.102$). Distributional analysis of the treatment effects suggests that the program significantly improves problem behavior among children with above-average problem load at the 75^{th} and 90^{th} percentiles of the outcome distribution (Table 4).

The reduction in behavioral difficulties appears to be driven by a reduction in externalizing behaviors (column 3, 0.11 SD reduction, $p = 0.116$), with significant impacts at the top half of the externalizing problems distribution (Table 4). These reductions are consistent with the program's emphasis on structured, adult-supervised small-group play, which promotes rule-following and turn-taking. This is in line with evidence that play-based activities can strengthen children's self-regulation skills (Diamond and Lee, 2011; Healey and Healey, 2019), and with a broader literature showing that weaknesses in self-regulation are closely linked to externalizing difficulties such as hyperactivity and inattention (Sher-Censor et al., 2016; Coelho et al., 2023).

Analysis of treatment effects on boys and girls separately shows that the reductions in problem behaviors are concentrated among boys; *Lively Minds* reduces problem behaviors among boys by just over a fifth of a standard deviation (Table 5, Panel B, column 3, $p = 0.004$). This is consistent with broader evidence that boys tend to exhibit higher levels of externalizing problem behaviors than girls in early childhood—in our control group, on average, girls' problem behaviors score is 0.11 SD lower than boys'—and that the effect of *Lively Minds* is concentrated at the higher end of the problem score distribution (Keenan and Shaw, 1997).

In contrast to the findings for cognitive development, results in Table 8 suggest that improvements in problem behaviors are concentrated among children of PMs. Impacts among this group are large and significantly different from the small and statistically insignificant impacts on problem behaviors among children of non-PMs. Column 2 shows that among children of PMs, the program reduced behavioral difficulties by a third of a SD (column 2, Panel A, $p = 0.005$). This is a sizeable effect; much larger than the average effect of 0.1-0.15 SD reported in meta-analysis of preschool interventions (Holla et al.,

2021). This effect is driven by a similar size, significant reductions in both internalizing and externalizing behaviors.

This finding suggests that the program's benefits for children's socioemotional development are likely driven by improvements in the home environment and caregiving practices among PMs, stemming from gains in their ECD knowledge and behaviors. The absence of effects for children of non-PMs implies that changes occurring in preschools as a result of the program, while beneficial for cognitive outcomes, are insufficient on their own to improve children's socioemotional development in the absence of complementary improvements in the home environment.

4.3 Robustness

Selection of controls. Our main specifications follow our pre-specified approach to selection of controls, as described in Section 3. As a robustness check, we also estimate impacts on the cognitive aggregate score as well as the emotional awareness and behavioral difficulties scores, selecting controls using the post-double selection LASSO method (Belloni et al., 2014). The treatment effect estimates, shown in Table A10, are similar in magnitude and significance to the main estimates reported above.

Measurement error. We re-estimate the main specifications using cognitive scores constructed with Item Response Theory (IRT) to assess the sensitivity of our results to measurement error in the assessment scores. As discussed in Section 3 and Appendix B, IRT explicitly models measurement error and allows items to differ in difficulty and discrimination, rather than imposing equal weights across all items as in simple averages. Table A11 shows that the estimated treatment effects on the aggregate cognitive score, as well as on emergent numeracy and executive function, are very similar in magnitude and statistical significance to our main estimates using raw scores. Correcting for measurement error using IRT leads to a somewhat larger estimated effect on emergent literacy (from 0.049 to 0.075 SD, $p = 0.096$) and increases the significance of the treatment effect on externalizing behaviors. Table A12 shows the effects on the socioemotional outcomes.

Measurement invariance. A concern in interpreting treatment effects is that the intervention may affect how some assessment items function rather than the underlying skills themselves, which could bias comparisons between treated and control groups. We therefore test for measurement invariance for all latent constructs using standard multi-group factor models. We begin with a configural model that

allows all parameters to vary by group, and then sequentially test metric, scalar, and strict invariance by imposing equality restrictions on factor loadings, item intercepts, and residual variances, respectively, comparing model fit at each step. In practice, full scalar or strict invariance rarely holds in applied early-childhood settings, so we allow for partial invariance, freeing a small number of parameters where score tests indicate the largest violations while keeping the remaining items constrained to anchor the scale. Across our main constructs (cognitive skills, socioemotional skills, parental knowledge, and parenting behaviors), we find that while full invariance is often rejected, models with partial metric or partial scalar invariance fit the data as well as fully unconstrained models. For constructs where only partial invariance holds, we verify that our results are robust to using factor scores derived from the partially invariant model. As shown in Appendix Table B22, most estimated impacts are nearly identical or larger and more statistically significant than the main results, which use a simple average of raw item scores for measures of child development and IRT scores constructed assuming full measurement invariance. See Appendix B for a more in-depth discussion and analysis of this issue.

5 Opportunity costs and household spillovers

A common feature of early childhood and related social programs in low-resource LMIC settings is reliance on low-paid (or unpaid), often marginalised women as frontline implementers. A prominent example is the community health worker (CHW) model: in many settings, CHW cadres are predominantly female and are tasked with delivering a broad portfolio of services, often with limited compensation and high workload, raising concerns about exploitation and burnout (Jain et al., 2022; Astale et al., 2023; Ballard et al., 2023). In India, for instance, the all-female ASHA workforce has been widely debated in relation to incentive-based pay, stress, and excessive workload (Shrivastava et al., 2023; Dhaliwal et al., 2025). Pakistan's Lady Health Worker program provides another example of a large female cadre responsible for multiple maternal and child health tasks in underserved communities (Jalal, 2011). The *Lively Minds* program similarly depends on unpaid, low-educated mothers who already shoulder substantial domestic and care responsibilities. While such models can be cost-effective, a key welfare and scalability question is whether they impose hidden costs on participating women or their families. In spite of the critical role that such women play in the success of the ongoing global effort to improve early childhood provision, there is little evidence on how participation in the delivery of ECD programs affects them (Ballard et al., 2023; Sullivan et al., 2024).

In standard economic models of time allocation, volunteering is not "free": any additional program-related time must be accommodated through reductions in paid work, leisure, or other household activities (Becker, 1965; Gronau, 1977). This concern is particularly salient in low-income settings where women already face high unpaid care burdens and may be "time poor" (Bardasi and Wodon, 2010; Dinkelman and Ngai, 2022). If participation in the *Lively Minds* program meaningfully tightened these constraints, we would expect this to be reflected in mothers' psychosocial well-being.

Time requirements and psychosocial costs. In practice, the program's direct time requirement was modest—typically around 1–2 hours per week at the preschool—and predictable, making it easier to absorb through small adjustments in daily routines rather than major changes in labor supply. Moreover, many women's economic activities in this rural setting are informal, seasonal, and flexible, which likely reduced the likelihood of large crowd-out effects.

Consistent with this, we find no evidence that the program adversely affected participating mothers' mental health or self-esteem (Table 9).[11] A plausible interpretation is that the limited time commitment did not materially increase net time pressure. In addition, any incremental burden may have been offset by potential psychosocial benefits from volunteering, social interaction, and a greater sense of purpose or community engagement—channels that have been linked to better well-being in other contexts (Piliavin and Siegl, 2007; Hussam et al., 2022). *Lively Minds* program training and monthly parenting workshops also explicitly emphasized maternal well-being, which may have further mitigated potential stress.[12]

Spillovers on siblings. We next examine whether the program generated spillovers within the household. This is in line with evidence that parent-focused interventions often generate broader familial impacts (Carneiro et al., 2023). Spillovers could be positive if (i) target children transmit what they learned at preschool to siblings, or (ii) participating mothers apply new knowledge and practices to all children in the household. Conversely, spillovers could be negative if mothers reallocated attention toward the target child at the expense of siblings.

To examine these possibilities, we estimate impacts on one younger and one older sibling of each target child (see Section 3) and report them in Table 10. The sample includes 749 younger and 1,083 older siblings. We find no evidence of program impacts on older siblings. For younger siblings, we

[11] See Section 3 for details on how these are measured.

[12] In fact, there is some suggestive evidence that the program may have had positive impacts on the mental health and self-esteem of participating mothers (PMs) in analyses that draw on an augmented booster sample of PMs.

observe a positive and economically meaningful impact, but the effect is imprecisely estimated (column 6, $p = 0.244$) due to the small sample (155 siblings). Overall, the results offer no evidence of indirect spillovers from children's participation, but improvements in parenting among PMs may benefit younger siblings. Importantly, we find no evidence of adverse effects on any children in the household.

6 Cost-effectiveness, scalability, and sustainability

Our results suggest that the *Lively Minds* program offers a promising approach to improving preschool learning environments in highly resource-constrained, poor, and rural settings in LLMICs. They demonstrate that it is feasible to train marginalized mothers with little or no formal education to support the delivery of structured, play-based learning in public preschools, without evidence of harm to their psychosocial well-being. Importantly. engaging mothers rather than, for example, paraprofessionals has the advantage of also generating additional benefits for the development of their own children. More broadly, the findings suggest that effective ECD inputs can be embedded within low-quality, low-capacity public education systems rather than relying on parallel delivery structures.

Whether this model is a viable policy option, however, depends not only on its effectiveness but also on its cost-effectiveness and scalability—questions we focus on in this section.

6.1 Cost effectiveness of implementation for the RCT

Implementation of the program during the RCT was relatively low-cost, as can be seen in Panel A of Table 11. Total program costs were $225,917 in 2025 USD values, which corresponds to $47 per child in 2025 USD values ($N = 4,800$), equivalent to $37 per child in implementation-year dollars.[13] The cost structure during implementation for the RCT reflects a model in which most resources are devoted to staffing and field logistics rather than expensive classroom inputs.

To benchmark cost-effectiveness, we translate the impacts on cognitive skills into equivalent years of schooling (EYOS) following the approach set out in Evans and Yuan (2019) and using their estimate of a 1 SD gain in learning corresponding to 4.4 EYOS. Applying this conversion to our aggregate cognitive

[13]The largest cost component of the RCT was program administration and staff (e.g., salaries, fringe, and honorariums: $129,533.86; 57% of total), followed by transportation and per diems (e.g., vehicle rental, fuel, meals, lodging, etc.: $45,006.27; 20% of total), and office costs (e.g., rent, utilities, office supplies, etc.: $29,297.44; 13% of total). Direct implementation and materials (e.g., program supplies, intervention materials, and translation: $11,045.03), targeting/visibility (e.g., recruiting expenses, promotions, and items for promotions: $5,787.88), and training (e.g., staff and parents training: $5,246.14) were comparatively small shares of the budget (about 10% of the total).

effect (0.13 SD) implies an EYOS gain of 0.572 years per child. Using our program's implementation-year cost of $37 per child, this corresponds to 1.55 EYOS per $100 spent: that is, roughly 1.6 additional years of business-as-usual schooling per $100 of program expenditure.

Importantly, the RCT delivery model was necessarily more costly on a per-child basis than routine implementation. *Lively Minds* program is typically implemented at the district level, with fixed costs spread over a "saturated" set of all preschools in the district. For the trial, the program had to instead be randomised within districts at the preschool level, incurring many of the usual district-level fixed costs while reaching less than half of the preschools that they would reach in routine implementation. Moreover, preliminary evidence based on cost and implementation data from the ongoing scale-up of the program suggests significant scope for leveraging economies of scale to reduce costs as coverage expands. We now turn to this scale-up phase.

6.2 Scalability and sustainability

As is increasingly highlighted in the literature scaling even the most promising programs is fraught with challenges. There is a significant risk of "voltage drop" in effectiveness for a variety of reasons, including, among others, lack of suitably skilled workforce, weak incentives for maintaining quality in delivery, and the possibility of encountering diseconomies of scale (List et al., 2021). This risk is particularly salient in the delivery of ECD programs because of the critical importance of maintaining implementation quality, and growing evidence on effective pilots losing impact when taken to scale (Andrew et al., 2018; Britto et al., 2018).

These concerns raise several challenges for the sustainability of the impacts we document from the *Lively Minds* program at scale. First, while the program is embedded in existing government infrastructure, the implementation model evaluated in the RCT relied on substantial input from Lively Minds NGO to support setup and monitor implementation. A sustainable model would therefore require the government to build this capacity internally. Second, the program relies on a volunteer-based model. Its scalability depends on whether high levels of parental interest and commitment can be sustained across a broader set of rural communities, and whether turnover among participating mothers can be effectively managed. Third, given the context of a shrinking education budget in Ghana (World Bank and UNESCO, 2024) and per-pupil expenditure at the KG level of around $90 in 2025, an important question is whether further cost reductions can be achieved through economies of scale.[14]

[14]The per pupil expenditure is calculated based on figures provided in Ministry of Education, Republic of Ghana (2025),

Our study is unusual in that we can go beyond speculation about these scalability concerns. The findings from this evaluation motivated the Government of Ghana to expand the coverage of the *Lively Minds* program to 64 high-need rural districts in Northern Ghana, reaching 532,413 children — over 40% of all children enrolled in public preschools in Ghana — through a gradual scale-up between 2021 and 2024.

We draw on information about the adjustments made to the implementation model for scaling, as well as cost, implementation, and fidelity data from this scale-up, to shed light on the key sustainability and scalability issues outlined above.

6.2.1 Implementation adjustments

The main adaptations made to further embed the program within the Ghanaian education system and enable delivery at scale included: (1) strengthening quality assurance systems within the GES; (2) reducing the role of Lively Minds NGO in set-up and monitoring; and (3) increasing non-financial incentives for pre-school teachers and participating parents.

Quality assurance systems: A central GES program management unit was established to provide oversight to all districts implementing the program. The program management unit holds regular quality assurance calls with a focal person from each district, conducts supervisory visits and helps solve challenges. Districts are expected to meet termly performance targets.

Integration of program monitoring tools and procedures with GES monitoring systems: As part of the scale-up, the program's monitoring arrangements were fully embedded within routine GES supervisory processes. Play Scheme monitoring is now carried out by GES circuit supervisors using an electronic supervision form that has been integrated into the standard set of tools used during regular school visits. In addition, district teams were trained to meet twice per term to review supervision data, identify implementation risks, and agree on corrective actions.

Training facilitation. In the implementation model evaluated in the RCT, training of PMs was led by pre-school teachers but supervised primarily by Lively Minds staff. Under scale-up, supervision is shifted to district GES teams and high-performing KG teachers from districts already implementing the program, with a smaller supporting role for *Lively Minds* (e.g., one staff member per district team). Monthly top-up trainings are replaced with twice-per-term workshops run by district teams and peer trainers for pre-school teachers and headteachers.

which reports the KG budget for 2025 to be GHS 1,272,097,816 and total enrollment in public KGs of 1,233,676 pupils.

Incentives for implementers: To strengthen the scale-up model, GES introduced several new non-financial incentive mechanisms for frontline implementers. This included official certification for participating mothers who complete the training, which is salient given that many have no formal qualifications. In addition, teachers who meet termly performance targets become eligible for professional development points administered by the National Training Council. Finally, high-performing teachers are recognized through being offered the opportunity to train to become peer facilitators to support training-of-trainers in new districts and to deliver refresher sessions.

Following these adjustments, involvement of the *Lively Minds* program in implementation shifted primarily to providing capacity-building support to the district teams administering the program. Specifically, they are offered a package of support for two years after the adoption of the program, which includes coaching, termly capacity building workshops, and advice on solving challenges.

6.2.2 Compliance and fidelity at scale

We draw on monitoring data from GES for the first two years of implementation in 46 out of the 64 districts included in the scale-up, covering 2,543 preschools across eight regions over five preschool terms between 2022 and 2024 (the exact period varies by district depending on when rollout began).[15] This data allows us to assess whether parental interest and commitment to the program, as well as overall program compliance, were sustained at scale.

Parental engagement. The set-up phase for the PMs began in 2022, when mothers were recruited and trained through the Parenting Workshops (PW) to run the Play Scheme (PS) sessions. In 2023–24, another round of recruitment and training took place, and many PMs from 2022 had the opportunity to re-enroll.

Across the scale-up, the program appears to attract and retain a large pool of PMs, which is summarized in Table 12. A total of 96,339 women signed up to participate across the 2,543 preschools (average of 37.9 mothers per preschool). Of these, 91,761 completed the training (average of 36.1 mothers per preschool) completed the set-up training, corresponding to a very high training completion rate of 95%.

Importantly, participation also appears durable. Two years after initial enrollment, preschools report that an average of 30.9 existing PMs re-enrolled, implying a mean re-enrollment rate of 86%. While

[15] Each academic year in Ghana has three terms. We do not yet observe comparable data for the 18 districts where implementation started after 2023.

there is meaningful variation across schools, the overall picture is one of persistent engagement over a multi-year horizon. Taken together, the sign-up, training completion, and re-enrollment figures suggest that, even with reduced reliance on Lively Minds NGO and delivery through government systems at scale, it remains feasible to mobilize and retain a large pool of participating mothers—an essential condition for sustaining the intensity of classroom support needed for the program to continue generating impacts at scale.

Compliance. We next examine whether the strong parental engagement translated into consistent implementation of PW and PS. Using unannounced GES monitoring visits data during the scale-up, Panel A of Figure 1 documents whether PS and PW sessions were running at the time of the visit. For PS, GES officials classified a session as running if more than five PMs were present and at least one teacher was present. For PW, a workshop was classified as successfully running if more than 75% of the expected PMs were present. The results suggest high and stable implementation coverage overall. PS sessions were running in a large majority of observed visits across terms, with an early level in the mid-to-high 90% range, a small dip around Term 4, and a recovery by Terms 5–6. PW sessions start below PS in Term 2, rise by Term 3, remain high through Term 4, and then fluctuate modestly while staying close to PS in later terms.

For comparison, RCT monitoring visits typically recorded 2 teachers present and an average of 8 PMs during PS sessions. In addition, each school received on average 4-5 monitoring visits during the monthly PWs over the academic year. In 88% of these visits, the workshop was taking place; in most cases, both teachers were present, and average attendance was 30 mothers per workshop.

Fidelity. Panel B of Figure 1 focuses on whether established quality thresholds were met in PS sessions, conditional on sessions being observed as running. Among those sessions, GES officials rated a very high share of both PS and PW sessions as satisfactory across all terms. Satisfactory quality for both components appears to be around the 90%+ range throughout, rising slightly and peaking around term 4 before tapering modestly thereafter. The close tracking of PS and PW quality over time suggests that when schools and PMs were able to run sessions, implementation quality was consistently strong. Overall, these patterns provide reassuring evidence that the scale-up maintained not only session continuity but also a good standard of delivery for both play-based classroom activities and parenting engagement.

6.3 Costs at scale

Building on suggestive evidence from GES monitoring data of sustained parental engagement, high compliance, and strong implementation fidelity, we next examine how the implementation adjustments described earlier, such as greater reliance on GES systems, reduced NGO involvement, and expansion to all preschools within participating districts, as well as overall scale-up of operations, impacted program costs.

Panel B, Table 11 shows that total scale-up expenditures were around $9.26 million (in 2025 USD values) covering 532,413 children. This implies an average cost per child of $17.4 in 2025 USD. This is nearly 3 times less than the per-child cost of implementation during the RCT ($47 in 2025 dollars). The reduction is driven primarily by large declines in per-child "fixed" and overhead-type costs. Program administration and staff fell from $26.99 per child in the RCT to $9.15 at scale, and office costs fell sharply from $6.10 to $0.64 per child. Transportation and per diems also declined materially on a per-child basis, from $9.38 to $5.85. Smaller per-child reductions occur in implementation and materials ($2.30 to $1.74). Two categories that were non-trivial in the RCT effectively disappear in the scale-up budget: targeting/visibility ($1.21 to $0) and training ($1.09 to roughly $0 per child). Overall, while the composition of spending remains broadly similar, with administration/staff and transport remaining the two largest categories by share, the sharp drop in cost per child is consistent with economies of scale and the dilution of central administration and overheads across a much larger beneficiary base.

To benchmark these costs against other education investments, we follow Angrist et al. (2025) and report cost-effectiveness in learning-adjusted years of schooling (LAYS), which is a standardized "learning per dollar" metric designed to make education interventions comparable across countries and implementation contexts. This benchmark complements the raw cost-per-child comparisons by placing this program on a common cross-intervention cost-effectiveness scale. These calculations are made on the assumption that, given the high level of compliance and fidelity suggested by GES monitoring data, we can expect RCT impacts on cognitive development of 0.13 SD to be sustained.

Following Angrist et al. (2025), we compute micro-LAYS as $(\beta/\delta_h) \times t$, where β is our cognitive impact in SD units, δ_h is the benchmark annual learning rate in a high-performing system (0.8 SD/year), and t is the duration over which impacts are assumed to persist (we set $t = 1$). LAYS per $100 is then computed as LAYS gained per child divided by per-child cost, multiplied by 100. Using $\beta = 0.13$ SD and a per-child scale-up cost of $17.4, we obtain LAYS $= (0.13/0.8) \times 1 = 0.1625$, which implies

$(0.1625/17.4) \times 100 = 0.93$ LAYS per \$100. Thus, the program delivers close to 1 learning-adjusted year of schooling per \$100 of program spending during scale-up. This places the program in the top half of the "cost-effective" (positive-impact) interventions reported in Angrist et al. (2025), which is comparable to the Teaching at the Right Level (TaRL) program under government implementation in India.

Overall, the evidence on parental engagement, compliance, fidelity, and costs at scale is encouraging, suggesting that the core elements of the *Lively Minds* model have been preserved under government-led delivery thus far. Nevertheless, scale-up is still ongoing, and it remains to be seen whether these positive implementation patterns will be sustained as the program continues to expand and become more established across additional districts and cohorts. Equally important, the current roll-out has benefited from a two-year package of technical and capacity-building support from Lively Minds NGO; it is therefore an open question whether similarly high levels of participation, fidelity, and cost efficiency can be maintained once this support is fully withdrawn and responsibility for training, supervision, and quality assurance rests entirely with routine GES systems. Continued monitoring will be critical to assess the durability of these gains over the longer run.

7 Conclusion

Many LLMICs have expanded access to pre-primary schooling in recent years, but face financial and administrative constraints in improving what children actually learn in these settings. Returns to investing early are well established: skill formation is dynamic, early investments raise the productivity of later investments, and early deficits are costly to remediate, implying potentially high returns to improving the quality of early learning environments (Heckman, 2012; Gertler et al., 2014). Yet the binding policy question is not whether to expand access, but how to improve the quality of existing preschools at scale in systems where teachers are overstretched, learning materials are scarce, education budgets are shrinking, and many caregivers have limited education themselves.

This paper studies a government-led approach to this problem in rural Ghana. In partnership with Ghana's Ministry of Education and the NGO Lively Minds, we evaluate a program using a cluster-RCT that recruits mothers from the local community, trains them using simple, scripted materials that do not require literacy, and embeds them as regular volunteers in public preschools to deliver structured, play-based learning activities in small groups. An important innovation is that the program uses mothers,

rather than hired paraprofessionals or teachers, as the marginal classroom input, so that implementation itself can build parenting human capital and potentially improve the home environment concurrently.

We observe economically meaningful gains in children's cognitive development. The program increases an aggregate cognitive index by 0.13 SD, with impacts concentrated in emergent numeracy and executive function. Socioemotional impacts, however, are more selective: reductions in behavioral difficulties are concentrated among boys only. We also find that cognitive gains accrue broadly to all children exposed to improved classroom activities, suggesting that low-educated mothers can effectively augment teacher capacity when provided with structured training. On the other hand, reductions in behavioral difficulties are driven entirely by children of participating mothers, consistent with large improvements in mothers' knowledge of child development, greater engagement in play activities at home, and increases in observed positive parenting behaviors. This implies that engaging mothers as active contributors, rather than relying exclusively on classroom-based inputs, generates complementary benefits through the home environment channel, particularly for socioemotional outcomes. Importantly, we find no adverse effects on mothers' mental health or self-esteem, addressing concerns about the hidden costs of volunteer-based models in settings where women already face substantial care burdens.

Finally, we provide novel evidence on scalability. Post-trial, the Government of Ghana expanded the model across 46 rural districts. While the scale-up data do not include child outcomes, administrative records from 2,543 kindergartens show high volunteer training completion and re-enrollment of mothers, and unannounced monitoring visits from the government indicate that both parenting workshops and play scheme sessions are typically running and are rated satisfactory when observed. The program costs fell sharply at scale to roughly $17.4 per child (compared to $37 during the trial) due to substantial economies of scale and learning.

Two features of the *Lively Minds* program help rationalize why this model performs well under severe implementation constraints. First, the "production technology" is intentionally simple relative to standard parenting programs: it does not require literacy, relies on repeatable activity scripts, and is delivered to groups of children rather than through intensive one-to-one coaching at home. Second, and more unusually, the main play-activity input is provided by low-educated, often illiterate mothers from the community. Given limited prior evidence on whether parents with very low schooling in poor rural LLMIC settings can effectively support children's learning in formal classroom environments, our results suggest that — with appropriate structure and support — they can. A likely reason for this success

is that the program was developed iteratively over more than a decade in Northern Ghana, adapting the model to local constraints rather than transplanting a first-best design into weak systems.

Whether these results generalize to other low-capacity countries or contexts remains an open question. The model's reliance on volunteer mothers may be more feasible in contexts where women have flexible work schedules, strong community ties, and intrinsic motivation to support children's education - conditions that may not hold uniformly across LLMICs. More work is also needed on equilibrium and welfare effects, including how volunteer models affect women's labor supply and intrahousehold allocations over longer horizons, and whether communities experience volunteer fatigue or selection in volunteer supply as programs mature.

References

Amadu, S., O. Attanasio, B. Caeyers, L. Cardona Sosa, S. Cattan, S. Krutikova, P. Leighton, L. Masselus, and M. Yakubu (2018). Improving early childhood development in rural ghana through scalable low-cost community-run play schemes: Baseline report. IFS Report R144.1, Institute for Fiscal Studies.

Amadu, S., O. Attanasio, B. Caeyers, L. Cardona Sosa, S. Cattan, S. Krutikova, P. Leighton, and M. Yakubu (2020). Improving early childhood development in rural ghana through scalable community-run play schemes: Programme impact evaluation report. Technical report, Institute for Fiscal Studies, London, UK. Accessed: 30 December 2025.

Andrew, A., O. Attanasio, E. Fitzsimons, S. Grantham-McGregor, C. Meghir, and M. Rubio-Codina (2018). Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in colombia. *PLOS Medicine 15*(4), e1002556.

Andrew, A., O. P. Attanasio, R. Bernal, L. C. Sosa, S. Krutikova, and M. Rubio-Codina (2024). Preschool quality and child development. *Journal of Political Economy 132*(7), 2304–2345.

Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, H. Rogers, and S. Sabarwal (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics 172*, 103382.

Angrist, N., S. Kabay, D. Karlan, L. Lau, and K. Wong (2025). Human capital at home: Evidence

from a randomized evaluation in the philippines. NBER Working Paper 33574, National Bureau of Economic Research.

Astale, T., T. Abebe, and G. Mitike (2023). Workload and emerging challenges of community health workers in low-and middle-income countries: A mixed-methods systematic review. *PLoS One 18*(3), e0282717.

Attanasio, O., S. Cattan, and C. Meghir (2022). Early childhood development, human capital, and poverty. *Annual Review of Economics 14*(1), 853–892.

Ballard, M., A. Olaniran, M. M. Iberico, A. Rogers, A. Thapa, J. Cook, Z. Aranda, M. French, H. E. Olsen, J. Haughton, et al. (2023). Labour conditions in dual-cadre community health worker programmes: a systematic review. *The Lancet Global Health 11*(10), e1598–e1608.

Bardasi, E. and Q. Wodon (2010). Working long hours and having no choice: Time poverty in guinea. *Feminist Economics 16*(3), 45–78.

Bassi, M., B. Besbas, L. I. D. Diaz, S. Ravindran, and A. Reynoso (2025). From access to achievement: The primary school-age impacts of an at-scale preschool construction program in highly deprived communities. *NBER working paper*.

Becker, G. S. (1965). A theory of the allocation of time. *The Economic Journal 75*(299), 493–517.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies 81*(2), 608–650.

Blimpo, M. P., P. Carneiro, P. Jervis, and T. Pugatch (2022). Improving access and quality in early childhood development programs: Experimental evidence from the gambia. *Economic Development and Cultural Change 70*(4), 1479–1529.

Bornstein, M. H., L. R. Cote, O. M. Haynes, C.-S. Hahn, and Y. Park (2010). Parenting knowledge: experiential and sociodemographic factors in european american mothers of young children. *Developmental Psychology 46*(6), 1677–1693.

Bouguen, A., D. Filmer, K. Macours, and S. Naudeau (2018). Preschool and parental response in a second best world: Evidence from a school construction experiment. *Journal of Human Resources 53*(2), 474–512.

Britto, P. R., M. Singh, T. Dua, R. Kaur, and A. K. Yousafzai (2018). What implementation evidence matters: scaling-up nurturing interventions that promote early childhood development. *Annals of the New York Academy of Sciences 1419*(1), 5–16.

Carneiro, P., E. Galasso, I. L. Garcia, P. Bedregal, and M. Cordero (2024). Impacts of a large-scale parenting program: Experimental evidence from chile. *Journal of Political Economy 132*(4), 1113–1161.

Carneiro, P., I. Rasul, and F. Salvati (2023). Families as drivers of inequality: Experimental evidence from an early childhood intervention. *Working Paper*.

Cascio, E. U. and D. W. Schanzenbach (2013). The impacts of expanding access to high-quality preschool education. *Brookings Papers on Economic Activity*, 127–178.

Coelho, V., J. Cadima, and A. I. Pinto (2023). Child functioning in inclusive preschools: Associations between self-regulation, engagement, prosociality, and hyperactivity. *Journal of Applied Developmental Psychology 86*, 101518.

Coffey, J. R. and E. S. Spelke (2024). Direct cognitive testing in the field at scale: Leveraging core knowledge to assess the development of young children in ghana. Unpublished manuscript / preprint.

Cooke, E., S. Hague, and A. McKay (2016, March). The ghana poverty and inequality report: Using the 6th ghana living standards survey.

Cunha, F., I. Elo, and J. Culhane (2022). Maternal subjective expectations about the technology of skill formation predict investments in children one year later. *Journal of Econometrics 231*(1), 3–32.

Currie, J. (2001). Early childhood education programs. *Journal of Economic perspectives 15*(2), 213–238.

Currie, J. and D. Almond (2011). Human capital development before age five. In *Handbook of Labor Economics*, Volume 4, pp. 1315–1486. Elsevier.

Dean, J. T. and S. Jayachandran (2019). Attending kindergarten improves cognitive but not socioemotional development: Evidence from india. Working paper.

Dhaliwal, B. K., M. Gupta, A. Nadda, S. Singh, A. Shet, K. Scott, A. Dutta, and S. Closser (2025). Urban community health workers in punjab, india: A qualitative study of ashas' roles in the health system. *PLOS Global Public Health 5*(8), e0004629.

Diamond, A. and K. Lee (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science 333*(6045), 959–964.

Dinkelman, T. and L. R. Ngai (2022). Time use and gender in africa in times of structural transformation. *Journal of Economic Perspectives 36*(1), 57–80.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review 109*(8), 2728–2765.

Evans, D. K. and F. Yuan (2019). Equivalent years of schooling: A metric to communicate learning gains in concrete terms. policy research working paper 8752. *World Bank policy research working paper 8752*.

Ganimian, A. J., K. Muralidharan, and C. R. Walters (2024). Augmenting state capacity for child development: Experimental evidence from india. *Journal of Political Economy 132*(5), 1565–1602.

Gertler, P., J. Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, and S. Grantham-McGregor (2014). Labor market returns to an early childhood stimulation intervention in jamaica. *Science 344*(6187), 998–1001.

Ghana Statistical Service and UNICEF (2018). Ghana multiple indicator cluster survey 2017/18, survey findings report. Ghana Statistical Service in collaboration with Ministry of Health, Ministry of Education, and UNICEF.

Goodman, R. (1997). The strengths and difficulties questionnaire: a research note. *Journal of Child Psychology and Psychiatry 38*(5), 581–586.

Grantham-McGregor, S. M., C. A. Powell, S. P. Walker, and J. H. Himes (1991). Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the jamaican study. *The Lancet 338*(8758), 1–5.

Gronau, R. (1977). Leisure, home production, and work—the theory of the allocation of time revisited. *Journal of Political Economy 85*(6), 1099–1123.

Hamadani, J. D., F. Tofail, A. Hilaly, S. N. Huda, P. Engle, and S. M. Grantham-McGregor (2010). Use of family care indicators and their relationship with child development in bangladesh. *Journal of Health, Population and Nutrition 28*(1), 23–33.

Healey, D. and M. Healey (2019). Randomized controlled trial comparing the effectiveness of structured-play (engage) and behavior management (triple p) in reducing problem behaviors in preschoolers. *Scientific Reports 9*(1), 3497.

Heckman, J. J. (2012). Invest in early childhood development: Reduce deficits, strengthen the economy. *The Heckman Equation 7*(1-2), 1–2.

Heckman, J. J. and S. Mosso (2014). The economics of human development and social mobility. *Annual Review of Economics 6*(1), 689–733.

Hirsh-Pasek, K., R. M. Golinkoff, L. E. Berk, and D. G. Singer (2009). *A mandate for playful learning in preschool: Applying the scientific evidence*. Oxford University Press.

Holla, A., M. Bendini, L. Dinarte Diaz, and I. Trako (2021, June). Is investment in preprimary education too low? lessons from (quasi) experimental evidence across countries. Policy Research Working Paper 9723, World Bank.

Hoosen, N., E. L. Davids, P. J. de Vries, and M. Shung-King (2018). The strengths and difficulties questionnaire (sdq) in africa: a scoping review of its application and validation. *Child and Adolescent Psychiatry and Mental Health 12*(1), 6.

Hussam, R., E. M. Kelley, G. Lane, and F. Zahra (2022). The psychosocial value of employment: Evidence from a refugee camp. *American Economic Review 112*(11), 3694–3724.

Jain, M., Y. Caplan, B. M. Ramesh, H. Kemp, B. Hammer, S. Isac, J. Blanchard, V. Namasivayam, and S. K. Sgaier (2022). Improving community health worker compensation: a case study from india using quantitative projection modeling and incentive design principles. *Global Health: Science and Practice 10*(3).

Jalal, S. (2011). The lady health worker program in pakistan—a commentary. *European Journal of Public Health 21*(2), 143–144.

Jeong, J., H. O. Pitchik, and G. Fink (2021). Short-term, medium-term and long-term effects of early parenting interventions in low-and middle-income countries: a systematic review. *BMJ Global Health 6*(3).

Keenan, K. and D. Shaw (1997). Developmental and social influences on young girls' early problem behavior. *Psychological Bulletin 121*(1), 95.

Lim, I. E., I. Mahmud, M. U. Chowdhury, A. Kaiser, F. A. Bonny, S. Akanbonga, and M. T. Hasan (2023). Early childhood development (ecd) in ghana: Assessing the status and determinants of the literacy–numeracy, physical, social–emotional, and learning domains. *International Journal of Early Childhood*. Published online 28 Feb 2023.

List, J. A., D. Suskind, and L. H. Supplee (Eds.) (2021). *The Scale-Up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and What We Can Do About It*. London and New York: Routledge.

MacPhee, D. (1981). *Knowledge of Infant Development Inventory (KIDI)*. Unpublished manual, University of North Carolina.

Malapit, H. J. L. and A. R. Quisumbing (2015). What dimensions of women's empowerment in agriculture matter for nutrition in ghana? *Food Policy 52*, 54–63.

Martinez, S., S. Naudeau, and V. Pereira (2017). Preschool and child development under extreme poverty: Evidence from a randomized experiment in rural mozambique. Working Paper 8290, World Bank, Washington, DC.

McCoy, D. C., C. R. Sudfeld, D. C. Bellinger, A. Muhihi, G. Ashery, T. E. Weary, W. Fawzi, and G. Fink (2017). Development and validation of an early childhood development scale for use in low-resourced settings. *Population Health Metrics 15*(1), 3.

Ministry of Education, Republic of Ghana (2025). 2025 Programme Based Budget Estimates: Ministry of Education. Programme based budget estimates, Ministry of Education, Republic of Ghana.

National Council for Curriculum and Assessment (NaCCA) (2023). 2022 P4 NST Report. National standardised test report, National Council for Curriculum and Assessment. Findings from the 2022 National Standardised Test for Basic Four pupils, administered and analysed by NaCCA.

Özler, B., L. C. Fernald, P. Kariger, C. McConnell, M. Neuman, and E. Fraga (2018). Combining preschool teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics 133*, 448–467.

Piliavin, J. A. and E. Siegl (2007). Health benefits of volunteering in the wisconsin longitudinal study. *Journal of Health and Social Behavior 48*(4), 450–464.

Romano, J. P. and M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association 100*(469), 94–108.

Sher-Censor, E., T. Y. Khafi, and T. M. Yates (2016). Preschoolers' self-regulation moderates relations between mothers' representations and children's adjustment to school. *Developmental Psychology 52*(11), 1793.

Shrivastava, R., L. Sharma, M. Jolly, R. Ahuja, R. Sharma, J. A. Naslund, J. Agrawal, R. Shidhaye, S. Mehrotra, S. D. Hollon, et al. (2023). "we are everyone's ashas but who's there for us?" a qualitative exploration of perceptions of work stress and coping among rural frontline workers in madhya pradesh, india. *Social Science & Medicine 336*, 116234.

Sullivan, J., Z. Aranda, M. Adusumilli, A. Martens, and A. Jiménez (2024). Perceptions around occupational mental well-being of community health workers and an intervention package for its promotion: A mixed-methods study in rural chiapas, mexico. *SSM-Mental Health 6*, 100360.

The World Bank (2024). Ghana Accountability for Learning Outcomes Project (P165557) – Implementation Status & Results Report. Implementation status & results report, World Bank. Implementation Status & Results Report for the Ghana Accountability for Learning Outcomes Project as of December 25, 2024.

UNICEF (2019a). Ghana multiple indicator cluster survey 2017/18, survey findings report. Technical report, Ghana Statistical Service, Accra, Ghana. Co-published by UNICEF.

UNICEF (2019b). A world ready to learn: Prioritizing quality early childhood education. Global report, United Nations Children's Fund (UNICEF), New York, NY.

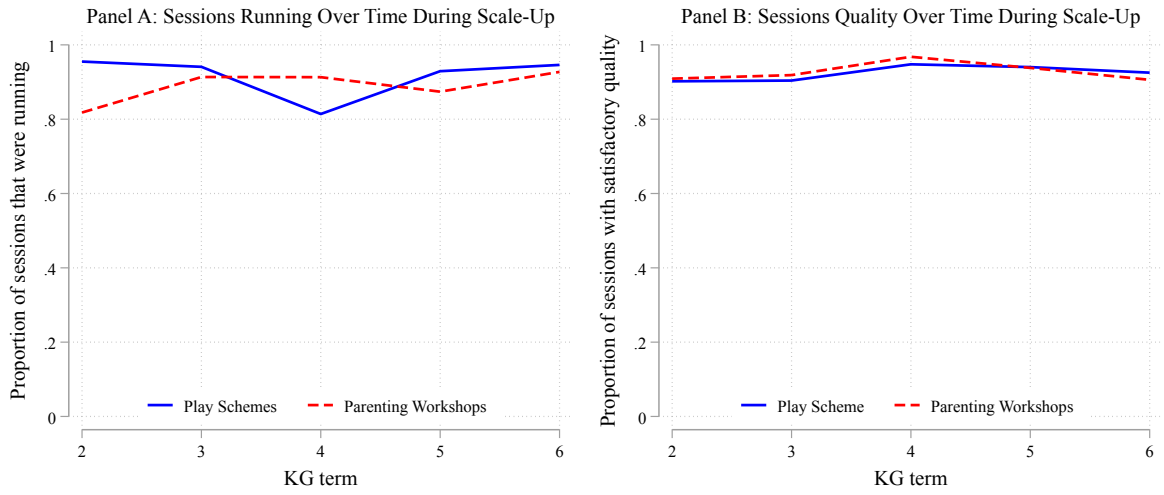
World Bank (2018a). The world development report 2018 - learning to realize education's promise. Wdr, World Bank, New York, NY.

World Bank (2018b). *World Development Report 2018:* Learning to Realize Education's Promise. World Bank.

World Bank and UNESCO (2024, January). Education Finance Watch 2024. Technical report, World Bank and UNESCO Institute for Statistics.

Main Figures and Tables

Figure 1: Implementation and compliance during scale-up



Note: Panel A shows the proportion of sessions observed to be running during unannounced visits. Panel B shows the proportion of observed sessions meeting a satisfactory quality threshold, from the 2nd to the 6th kindergarten term. Note that KG Term 1 is the set-up phase during which they recruit and train the PMs; hence, no session data are available.

Table 1: Teacher reports related to program compliance

	Parents Contribute	Parents support at least once a month	Any games for free play	Any play in small groups
	(1)	(2)	(3)	(4)
Treatment	4.088	0.382	0.291	0.052
	(1.671)	(0.081)	(0.089)	(0.041)
p-val	*0.001*	*0.000*	*0.002*	*0.213*
Control mean endline	2.154	0.372	0.526	0.885
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	156	156	156	156

Notes: The sample includes teachers who work in one of the pre-schools randomized into treatment or control at baseline and who were sampled to answer the teacher survey at endline. In Column (1), the coefficient is an odds ratio from an ordered logit model. In Columns (2)–(4), OLS estimates are presented. Across all columns, standard errors are clustered at the community level, and p-values are reported in italics. The model controls for strata fixed effects and an indicator for whether there is a clothing shop in the community. The outcomes and baseline outcomes are standardised to have mean 0 and standard deviation 1 in the control group. At the foot of each column we report the mean outcome among controls tracked to endline.

Table 2: Impacts on parental knowledge, play investments, and behaviors

	I. Knowledge		II. Play Investments		III. Behaviors	
	Child Development	Pre-school Quality	Play Materials	Any play activities	Positive Behaviors	Negative Behaviors
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Treatment among PMs	0.472	0.310	-0.095	0.180	0.260	0.082
	(0.141)	(0.115)	(0.146)	(0.051)	(0.120)	(0.135)
p-val	*0.001*	*0.009*	*0.515*	*0.001*	*0.033*	*0.544*
Adj. p-val	*0.016*	*0.062*	*0.766*	*0.012*	*0.116*	*0.766*
Panel B						
Treatment among non-PMs	-0.040	0.166	-0.241	-0.029	0.042	0.013
	(0.103)	(0.057)	(0.109)	(0.036)	(0.074)	(0.105)
p-val	*0.702*	*0.004*	*0.030*	*0.431*	*0.578*	*0.901*
Adj. p-val	*0.884*	*0.026*	*0.088*	*0.812*	*0.884*	*0.884*
Panel C						
Difference in treatment	0.512	0.143	0.146	0.209	0.219	0.069
	(0.132)	(0.109)	(0.130)	(0.046)	(0.099)	(0.142)
p-val	*0.000*	*0.194*	*0.265*	*0.000*	*0.031*	*0.629*
Adj. p-val	*0.009*	*0.359*	*0.374*	*0.002*	*0.078*	*0.631*
Control mean endline (non-PMs)	0.030	0.005	-0.118	0.317	0.009	0.028
Control mean endline (PMs)	0.073	0.308	-0.054	0.360	0.184	-0.011
Baseline outcome	Yes	No	Yes	Yes	No	No
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151	2151	2,151

Notes: All outcomes measured as IRT scores standardised to have mean 0, std 1 in control group. The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented where standard errors are clustered at the community level, and p-values are reported in italics. Column (1) uses an outcome from the KIDI scale. Column (2) uses a vignette-based instrument measuring parental knowledge about different aspects of pre-school quality. Columns (3) and (4) use outcomes from the Family Care Indicator. Columns (5) and (6) use outcomes from a surveyor-reported instrument on parenting behaviors toward the target child. Throughout we estimate fully interacted models that control for the baseline outcome (if available), strata fixed effects, child's age in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother signed up to participate in the play schemes (during the intervention in the treatment group, in the following year in the control group). Control means are reported for each subgroup at the bottom of the table.

Table 3: Impacts of intervention on children's cognitive outcomes

	Cognition Aggregate Score	Emergent Literacy	Emergent Numeracy	Executive Function
	(1)	(2)	(3)	(4)
Treatment	0.130	0.049	0.108	0.127
	(0.056)	(0.055)	(0.051)	(0.052)
p-val	*0.022*	*0.376*	*0.038*	*0.017*
Adj. p-val	-	*0.251*	*0.040*	*0.034*
Baseline outcome	0.408	0.313	0.381	0.231
	(0.023)	(0.024)	(0.022)	(0.022)
p-val	*0.000*	*0.000*	*0.000*	*0.000*
Control mean endline	0.000	0.000	0.000	0.000
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: The sample are those children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented where standard errors are clustered at the community level, and p-values are reported in italics. The outcome in Column 1 is the average of scores obtained on IDELA and Spelke tasks measuring Emergent Literacy, Emergent Numeracy, and Executive Function. Each score is measured as the percentage of items correct on each of these task. The outcomes in Columns 2 to 4 are the average of scores obtained on IDELA and Spelke tasks measuring each of the three subdomains. Throughout we control only for the baseline outcome measured as the average of scores on IDELA tasks measuring the same constructs, strata fixed effects, child's age measured in months of age at the time of the endline survey, a gender dummy, and a variable for the number of clothing stores in the community. The outcomes and baseline outcomes are standardised to have mean 0 and standard deviation 1 in the control group. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table 4: Distributional treatment effects: probability of scoring above selected control-group percentiles

	Dependent variable: Indicator for score above:				
	10^{th} pers	25^{th} pers	Median	75^{th} pers	90^{th} pers
	(1)	(2)	(3)	(4)	(5)
Cognition	0.015	0.042	0.044	0.076	0.016
	(0.013)	(0.014)	(0.023)	(0.029)	(0.018)
p-val	*0.253*	*0.004*	*0.055*	*0.012*	*0.370*
Emergent Literacy	0.005	0.014	0.024	0.019	0.012
	(0.014)	(0.020)	(0.021)	(0.026)	(0.016)
p-val	*0.739*	*0.483*	*0.261*	*0.462*	*0.468*
Emergent Numeracy	0.007	0.033	0.038	0.045	0.016
	(0.009)	(0.016)	(0.024)	(0.025)	(0.016)
p-val	*0.435*	*0.049*	*0.113*	*0.075*	*0.328*
Executive Function	0.024	0.051	0.017	0.031	0.025
	(0.013)	(0.017)	(0.023)	(0.028)	(0.018)
p-val	*0.068*	*0.004*	*0.475*	*0.268*	*0.172*
Emotional Awareness (aggregate)	-0.009	0.014	0.029	0.015	0.008
	(0.013)	(0.019)	(0.026)	(0.032)	(0.021)
p-val	*0.500*	*0.465*	*0.270*	*0.648*	*0.713*
Behavioral Difficulties	0.005	-0.021	0.002	-0.069	-0.054
	(0.016)	(0.029)	(0.027)	(0.024)	(0.016)
p-val	*0.738*	*0.476*	*0.928*	*0.004*	*0.001*
Externalising Behaviors	0.006	-0.010	-0.058	-0.045	-0.033
	(0.009)	(0.025)	(0.031)	(0.025)	(0.015)
p-val	*0.515*	*0.692*	*0.063*	*0.072*	*0.031*
Internalising Behaviors	0.008	0.019	0.019	-0.032	-0.053
	(0.010)	(0.024)	(0.032)	(0.027)	(0.022)
p-val	*0.425*	*0.445*	*0.543*	*0.250*	*0.018*
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes
Baseline outcome	Yes	Yes	No	Yes	No
Characteristics	Yes	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151	2151

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented; standard errors (in parentheses) are clustered at the community level. *P*-values are reported in italics. The dependent variable is an indicator for whether the measure of development is above a particular percentile of the endline score distribution. Models control for strata fixed effects, child age (in months) at endline, a gender dummy, and a variable for the number of clothing stores in the community; specifications also include baseline outcomes as indicated.

Table 5: Impacts on cognitive and socioemotional outcomes, by child's gender

	Cognition Aggregate Score	Emotional Awareness Aggregate Score	Behavioral Difficulties
	(1)	(2)	(3)
Treatment among boys	0.128	0.071	-0.211
	(0.060)	(0.079)	(0.070)
p-val	*0.036*	*0.372*	*0.004*
Treatment among girls	0.144	0.023	0.015
	(0.064)	(0.082)	(0.077)
p-val	*0.028*	*0.775*	*0.845*
Difference in treatment effect	0.016	-0.048	0.226
	(0.054)	(0.072)	(0.089)
p-val	*0.767*	*0.508*	*0.013*
Control mean endline – boys	0.064	0.063	0.022
Control mean endline – girls	0.032	-0.020	-0.085
Strata Fixed Effects	Yes	Yes	Yes
Baseline outcome	Yes	Yes	No
Characteristics	Yes	Yes	Yes
Observations	2151	2151	2151

Notes: The sample is those children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented where standard errors are clustered at the community level, and p-values are reported in italics. The outcome in Column 1 is the average of scores obtained on IDELA and Spelke tasks measuring Emergent Literacy, Emergent Numeracy, and Executive Function. The outcome in Column 2 is the average of scores obtained on IDELA and Spelke tasks measuring Emotional Awareness. The outcome in Column 3 is the average of scores obtained on the Strengths and Difficulties Questionnaire, measuring emotional problems, peer problems, conduct problems, and hyperactivity. Throughout, we estimate models that control for the baseline outcome, strata fixed effects, child's age measured in months of age at the time of the endline survey, a gender dummy, and a variable for the number of clothing stores in the community, and interactions of all these variables and the treatment dummy with an indicator for a female. At the foot of each Column, we report the mean outcome among controls in each of the subgroups.

Table 6: Impacts on children's cognitive outcomes, by PM status

	Cognition Aggregate Score	Emergent Literacy	Emergent Numeracy	Executive Function
	(1)	(2)	(3)	(4)
Panel A				
Treatment among PM children	0.134	-0.012	0.156	0.117
	(0.083)	(0.081)	(0.089)	(0.095)
p-val	*0.110*	*0.885*	*0.085*	*0.222*
Adj. p-val	-	*0.882*	*0.181*	*0.341*
Panel B				
Treatment among non-PM children	0.119	0.050	0.083	0.135
	(0.056)	(0.057)	(0.053)	(0.051)
p-val	*0.035*	*0.387*	*0.118*	*0.010*
Adj. p-val	-	*0.267*	*0.114*	*0.026*
Panel C				
Difference in treatment	0.015	-0.061	0.072	-0.019
	(0.072)	(0.074)	(0.087)	(0.087)
p-val	*0.834*	*0.409*	*0.409*	*0.830*
Adj. p-val	-	*0.719*	*0.719*	*0.812*
Control mean endline – non-PM children	0.024	-0.007	0.031	0.007
Control mean endline – PM children	0.149	0.081	0.160	0.094
Baseline outcome	Yes	Yes	Yes	Yes
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented with standard errors clustered at the community level, and p-values are reported in italics. The outcome in Column (1) is the average of scores obtained on IDELA and Spelke tasks measuring Emergent Literacy, Emergent Numeracy, and Executive Function. The outcomes in Columns (2) to (4) are the average of scores obtained on IDELA and Spelke tasks measuring each subdomain. Throughout we estimate fully interacted models that control for the baseline outcome, strata fixed effects, child's age in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother participates in the Lively Minds schemes. At the foot of each column we report the mean outcome among controls in the groups whose mother has signed up / not signed up to participate in the Lively Minds schemes in the following year.

Table 7: Impacts of intervention on children's socioemotional outcomes

		Behavioral difficulties		
	Emotional Awareness Aggregate Score	Aggregate Score	Externalising behaviors	Internalising behaviors
	(1)	(2)	(3)	(4)
Treatment	0.042	-0.098	-0.110	-0.061
	(0.073)	(0.059)	(0.061)	(0.060)
p-val	*0.560*	*0.102*	*0.074*	*0.314*
Adj. p-val	-	-	*0.116*	*0.238*
Baseline outcome	0.229	0.065	0.093	0.038
	(0.024)	(0.023)	(0.022)	(0.028)
p-val	*0.000*	*0.007*	*0.000*	*0.173*
Control mean endline	0.000	0.000	0.000	0.000
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: The sample are those children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented where standard errors are clustered at the community level, and p-values are reported in italics. The outcome in Column 1 is the average of scores obtained on IDELA and Spelke tasks measuring Emotional awareness. The outcome in Column 2 is the average of scores obtained on items of the Strengths and Difficulties Questionnaire (SDQ) that measure emotional problems, peer problems, conduct problems, and hyperactivity. Columns 3-4 disaggregate Column 2. The outcome in Column 3 is the average of scores obtained on items of the SDQ that measure emotional problems and peer problems. The outcome in Column 4 is the average of scores obtained on items of the SDQ that measure conduct problems and hyperactivity. Throughout we control strata fixed effects, child's age measured in months of age at the time of the endline survey, a gender dummy, and a variable for the number of clothing stores in the community. In Column 1, we also control for the average of scores obtained on items of the IDELA measuring Emotional Awareness at baseline. In Columns 2-4, we also control for the baseline outcome. The outcomes and baseline outcomes are standardised to have mean 0 and standard deviation 1 in the control group. At the foot of each Column we report the mean outcome among controls tracked to endline.

Table 8: Impacts on children's socioemotional outcomes, by PM status

		Behavioral difficulties		
	Emotional Awareness Aggregate Score	Aggregate Score	Externalising behaviors	Internalising behaviors
	(1)	(2)	(3)	(4)
Panel A				
Treatment among PM children	0.012	-0.332	-0.319	-0.257
	(0.088)	(0.115)	(0.116)	(0.119)
p-val	*0.892*	*0.005*	*0.007*	*0.034*
Adj. p-val	-	-	*0.036*	*0.032*
Panel B				
Treatment among non-PM children	0.038	-0.053	-0.078	-0.015
	(0.074)	(0.054)	(0.053)	(0.063)
p-val	*0.606*	*0.332*	*0.149*	*0.814*
Adj. p-val	-	-	*0.218*	*0.800*
Panel C				
Difference in treatment	-0.027	-0.279	-0.241	-0.242
	(0.072)	(0.104)	(0.096)	(0.118)
p-val	*0.713*	*0.009*	*0.014*	*0.044*
Adj. p-val	-	-	*0.045*	*0.045*
Control mean endline – non-PM children	0.013	-0.027	-0.029	-0.018
Control mean endline – PM children	0.060	-0.044	-0.066	-0.012
Baseline outcome	Yes	No	No	No
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented with standard errors clustered at the community level, and p-values are reported in italics. The outcome in Column (1) is the average of scores obtained on IDELA and Spelke tasks measuring Emotional Awareness. The outcome in Column (2) is the average of scores obtained on items of the Strengths and Difficulties Questionnaire (SDQ) that measure emotional problems, peer problems, conduct problems and hyperactivity. The outcome in Column (3) is the average of scores obtained on SDQ items measuring emotional problems and peer problems. The outcome in Column (4) is the average of scores obtained on SDQ items measuring conduct problems and hyperactivity. Throughout we estimate fully interacted models that control for the baseline outcome, strata fixed effects, child's age in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother participates in the Lively Minds schemes. At the foot of each column we report the mean outcome among controls in the groups whose mother has signed up / not signed up to participate in the Lively Minds schemes in the following year.

Table 9: Psychosocial impacts on the participating mothers

	SRQ-20 depressive symptoms score	Rosenberg self-esteem score
	(1)	(2)
Treatment among PMs	0.047	0.067
	(0.129)	(0.122)
p-val	*0.717*	*0.586*
Control mean endline – PMs	0.021	0.113
Baseline outcome	Yes	Yes
Strata Fixed Effects	Yes	Yes
Characteristics	Yes	Yes
Observations	422	422

Notes: In Column (1), the outcome is based on the SRQ-20 Depression Scale (mean of scores). In Column (2), the outcome is based on the Rosenberg Scale of Self-esteem (mean of scores). At the foot of each column, we report the mean outcome of the control group.

Table 10: Impacts on siblings of PM children

	Older siblings					**Younger siblings**
	Literacy Score (1)	Numeracy Score (2)	Geometric Intruder (3)	Vocabulary Assessment (4)	Point to Emotion (5)	CREDI Score (6)
Treatment among PM children	-0.067	0.028	0.016	-0.017	0.149	0.202
	(0.142)	(0.135)	(0.108)	(0.139)	(0.124)	(0.172)
p-val	*0.638*	*0.834*	*0.882*	*0.900*	*0.231*	*0.244*
Control mean endline (PM children)	0.167	0.213	0.094	0.069	-0.047	-0.045
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	242	242	242	242	242	155

Notes: The sample consists of the next older and next younger siblings of children who attended one of the preschools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented; standard errors (in parentheses) are clustered at the community level, and *p*-values are reported in italics. Outcomes in Columns (1)–(5) are latent scores from measures of the older sibling's literacy, numeracy, and tasks measuring geometric skills, vocabulary, and emotional awareness (the last three are measured using Spelke tasks). The outcome in Column (6) is the younger sibling's CREDI score, constructed using the CREDI tool's age standardisation procedure. All outcomes are standardised to have mean 0 and SD 1 in the control group. We estimate fully interacted models that control for the baseline outcome (for the older sibling, an overall score across literacy and numeracy at baseline), strata fixed effects, child age in months at endline, and the number of clothing stores in the community, as well as interactions of these controls and the treatment indicator with an indicator for whether the child's mother participates in the Lively Minds schemes. The younger sibling's age was not recorded and is therefore not included in the regression. Control means are reported separately for children whose mothers signed up / did not sign up to participate in the Lively Minds schemes.

Table 11: RCT and scale-up costs (in 2025 USD values)

Panel A: RCT implementation costs			
Cost category	Total cost	% of total	Cost per child
Program administration and staff costs	129,533.86	57.34	26.99
Implementation and program material costs	11,045.03	4.89	2.30
Transportation and per diems	45,006.27	19.92	9.38
Targeting / visibility costs	5,787.88	2.56	1.21
Training	5,246.14	2.32	1.09
Office costs	29,297.44	12.97	6.10
Total (N=4,800 treated children)	225,916.62	100.00	**47.07**
Panel B: Scale-up costs			
Cost category	Total cost	% of total	Cost per child
Program administration and staff costs	4,872,658.17	52.61	9.15
Implementation and program material costs	928,710.71	10.03	1.74
Transportation and per diems	3,116,933.67	33.65	5.85
Targeting / visibility costs	0.00	0.00	0.00
Training	2,602.74	0.03	0.00
Office costs	340,776.13	3.68	0.64
Total (N=532,413 treated children)	9,261,681.42	100.00	**17.40**

Note: Both cost analyses were conducted by Innovations for Poverty Action (IPA).

Table 12: Summary statistics on program participation by mothers

	Mean	SD	N	Min	Max	Sum
Number of PM who initially signed up/started the training	37.88	3.98	2,543	20.00	70.00	96,339
Number of PM who completed the training	36.08	4.46	2,543	20.00	67.00	91,761
Number of existing PMs re-enroling after 2 years	30.92	6.72	2,543	8.00	50.00	78,618
% of PMs completing set-up training after enrolment	0.95	0.07	2,543	0.57	1.00	2,424
% of PMs re-enroling after 2 years	0.86	0.15	2,543	0.22	1.00	2,177

Note: PM stands for Participating Mothers.

Parents in the Classroom: Strengthening Government Capacity to Deliver Early Childhood Education

Online Appendix

By Bet Caeyers, Lina Cardona-Sosa, Sarah Cattan, Sonya Krutikova, Abu Siddique

List of Tables

A1 Attrition: household and primary carer characteristics . . . 56
A2 Attrition: target child characteristics . . . 56
A3 Baseline balance: households and primary carer characteristics . . . 57
A4 Baseline balance: target child characteristics . . . 58
A5 Baseline balance using sample of PMs: households and primary carers characteristics . . . 59
A6 Baseline balance using sample of PMs: target child characteristics . . . 60
A7 Impacts on parental knowledge, play investments, and parenting behaviors (Mean of Scores) . . . 61
A8 Parental knowledge about preschool quality . . . 62
A9 Impacts on play materials and play activities . . . 63
A10 Impacts of intervention: Controls selected using post-double selection Lasso . . . 64
A11 Impacts on children's cognitive outcomes (IRT scores) . . . 64
A12 Impacts on children's socioemotional outcomes (IRT scores) . . . 65

List of Figures

A1 Parental knowledge about preschool quality picture . . . 55
B1 Instruments used to measure child development and parental investments . . . 66
B2 Instruments used to measure the target child's development . . . 78
B3 Correlation between items used to measure cognitive skill . . . 79
B4 Correlation between items used to measure emergent literacy . . . 80
B5 Correlation between items used to measure emergent numeracy . . . 80
B6 Correlation between items used to measure executive function . . . 80
B7 Correlation between items used to measure emotional awareness . . . 81

B8 Correlation between items used to measure emotional and behavioural problems 82
B9 Correlation between items used to measure externalising behaviours 83
B10 Correlation between items used to measure internalising behaviours 83
B11 Correlation between items used to measure knowledge of child development (KIDI meausures) 83
B12 Correlation between items used to measure pre-school quality 84
B13 Correlation between items used to measure play materials 85
B14 Correlation between items used to measure positive parenting behaviours 86
B15 Correlation between items used to measure negative parenting behaviours 87
B16 Correlation between items used to measure primary carer's depressive symptoms (SRQ-20) 88
B17 Correlation between items used to measure primary carer's self-esteem (Rosenberg scale) 89
B18 Tests of measurement invariance for cognitive measures 89
B19 Tests of measurement invariance for socio-emotional measures 90
B20 Tests of measurement invariance for measures of parental knowledge and parenting behaviours 91
B21 Tests of measurement invariance for measures of primary carer's mental health and self-esteem 92
B22 Impacts on measures allowing for partial measurement invariance, by PM 93

A Additional Figures and Tables

Figure A1: Parental knowledge about preschool quality picture

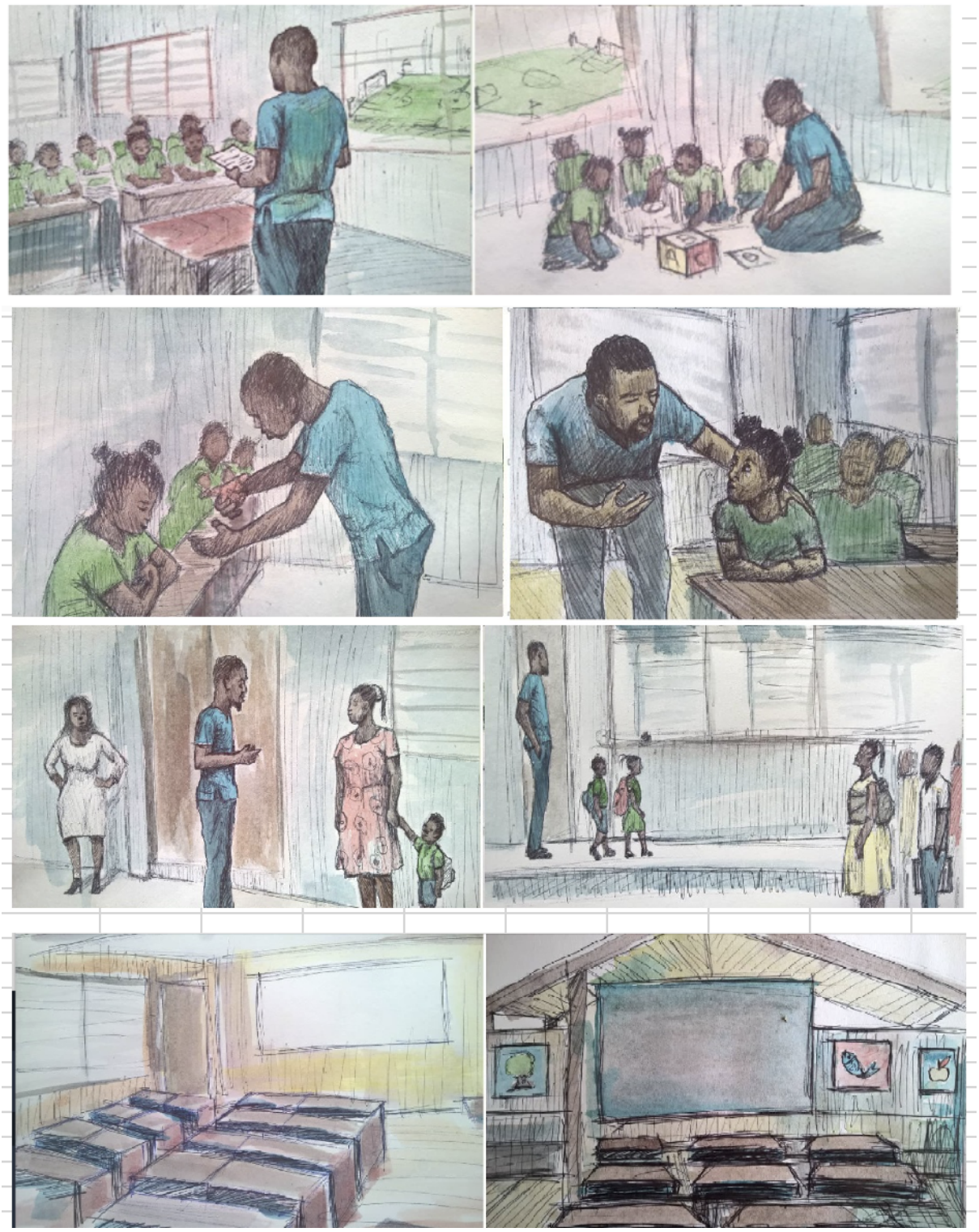


Table A1: Attrition: household and primary carer characteristics

	Endline sample		Attrited sample		p-value	p-value	
	Control	Treated	Control	Treated	attrit/endline	diff treated/control	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
A. Households							
Household size	8.965	10.321	9.992	10.367	0.002	0.250	2407
Number of children 16 or under	3.676	4.417	4.205	4.474	0.009	0.602	2407
Number of children 6 or under	1.332	1.617	1.536	1.639	0.045	0.667	2407
Wealth index (z-score)	-0.212	0.025	0.025	-0.025	0.067	0.361	2394
Main income source:							
Farming own land	0.359	0.446	0.435	0.438	0.044	0.842	2407
Wage work	0.328	0.274	0.276	0.284	0.665	0.221	2407
Livestock	0.047	0.033	0.035	0.033	0.036	0.075	2407
Profits from small enterprise	0.219	0.183	0.183	0.191	0.254	0.815	2407
B. Primary carers							
Age	36.648	36.435	36.092	36.833	0.687	0.744	2407
Target child's biological mother	0.656	0.790	0.789	0.763	0.000	0.949	2407
Some education	0.289	0.196	0.211	0.201	0.004	0.575	2407
Illiterate	0.859	0.926	0.915	0.923	0.003	0.589	2407
Raven's score	45.280	43.178	43.094	43.717	0.385	0.480	2407
Depression score	6.949	8.340	8.332	8.048	0.019	0.492	2407
Self-esteem score	18.637	17.944	18.105	17.928	0.000	0.014	2407

Notes: The sample is households and primary caregivers in the baseline sample. Columns (1) and (2) report means of baseline characteristics among observations in the endline sample, by treatment status. Columns (3) and (4) report the same means among observations that attrited at endline. We run a regression of each characteristic on an indicator for attrition (not being in the endline sample), a treatment indicator, and their interaction. Column (5) reports the p-value on the attrition indicator, and Column (6) reports the p-value on the interaction. Column (7) reports the full sample size.

Table A2: Attrition: target child characteristics

	Endline sample		Attrited sample		p-value	p-value	
	Control	Treated	Control	Treated	attrit/endline	diff treated/control	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Age (months)	55.311	56.325	56.209	56.226	0.104	0.487	2405
Gender	0.551	0.492	0.498	0.499	0.081	0.483	2407
Currently in school	0.766	0.781	0.779	0.780	0.870	0.734	2407
In the last 30 days:							
Any cough	0.512	0.549	0.567	0.522	0.975	0.281	2406
Any stomach pain	0.477	0.508	0.532	0.478	0.677	0.122	2406
Any high fever	0.340	0.340	0.359	0.321	0.651	0.522	2407
Any diarrhea	0.287	0.320	0.313	0.321	0.116	0.251	2395

Notes: The sample is the target children observed in the baseline sample. Columns (1) and (2) report means of baseline characteristics among observations in the endline sample, by treatment status. Columns (3) and (4) report the same means among observations that attrited at endline. We run a regression of each characteristic on an indicator for attrition (not being in the endline sample), a treatment indicator, and their interaction. Column (5) reports the p-value on the attrition indicator, and Column (6) reports the p-value on the interaction. Column (7) reports the full sample size.

Table A3: Baseline balance: households and primary carer characteristics

	Full sample	Control	Treatment	p-value	Adj. p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)
A. Households						
Household size	10.321	10.497	10.150	0.208	0.671	2151
	(6.369)	(6.756)	(5.967)			
Number of children 16 or under	4.417	4.573	4.264	0.045	0.228	2151
	(3.568)	(3.800)	(3.321)			
Number of children 6 or under	1.617	1.676	1.560	0.140	0.561	2151
	(1.825)	(1.893)	(1.755)			
Wealth index (z-score)	0.025	-0.025	0.074	0.283	0.729	2139
	(2.114)	(2.067)	(2.159)			
Main income source:						
Farming own land	0.446	0.447	0.445	0.933	0.990	2151
Wage work	0.274	0.274	0.274	0.998	0.994	2151
Livestock	0.033	0.034	0.031	0.721	0.970	2151
Profits from small enterprise	0.183	0.189	0.177	0.458	0.882	2151
B. Primary carers						
Age	36.435	36.948	35.935	0.050	0.251	2151
	(11.996)	(12.156)	(11.823)			
Target child's biological mother	0.790	0.774	0.806	0.063	0.291	2151
Some education	0.196	0.192	0.200	0.652	0.896	2151
Illiterate	0.926	0.930	0.922	0.466	0.882	2151
Raven's score	43.178	43.473	42.890	0.413	0.882	2151
	(16.524)	(17.090)	(15.956)			
Depression score (1–20)	8.340	8.189	8.486	0.193	0.593	2151
	(5.288)	(5.278)	(5.295)			
Self-esteem score	17.944	17.912	17.975	0.684	0.896	2151
	(3.584)	(3.656)	(3.514)			

Notes: The sample includes households and primary carers whose children attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. Columns (1)–(3) report baseline means for the full sample, control group, and treatment group (with standard deviations in parentheses underneath the means for non-binary variables). Columns (4) and (5) report the p-value and adjusted p-value for the difference in means between control and treatment; adjusted p-values use the Romano and Wolf (2005) procedure with 500 bootstrap samples. Column (6) reports the full sample size.

Table A4: Baseline balance: target child characteristics

	Full sample	Control	Treatment	p-value	Adj. p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)
A. Characteristics						
Age (months)	56.325	56.276	56.372	0.816	0.996	2151
	(9.499)	(9.481)	(9.521)			
Gender	0.492	0.496	0.489	0.754	0.996	2151
Currently in school	0.781	0.783	0.779	0.809	0.996	2151
In the last 30 days:						
Any cough	0.549	0.525	0.572	0.025	0.144	2150
Any stomach pain	0.508	0.481	0.535	0.011	0.068	2150
Any high fever	0.340	0.319	0.361	0.036	0.154	2151
Any diarrhea	0.320	0.321	0.320	0.975	0.996	2144
B. Cognitive outcomes						
Aggregate score	-0.009	0.000	-0.018	0.690	0.948	2151
	(1.040)	(1.000)	(1.078)			
Emergent literacy	-0.010	0.000	-0.020	0.650	0.948	2151
	(1.030)	(1.000)	(1.058)			
Emergent numeracy	0.017	0.000	0.033	0.462	0.918	2151
	(1.045)	(1.000)	(1.088)			
Executive function	-0.029	0.000	-0.057	0.202	0.687	2151
	(1.034)	(1.000)	(1.065)			
C. Socioemotional outcomes						
Emotional awareness	0.018	0.000	0.036	0.415	0.918	2151
	(1.038)	(1.000)	(1.073)			
Strengths and Difficulties:						
Emotional and behavioral problems	0.003	0.000	0.005	0.907	0.948	2151
	(0.987)	(1.000)	(0.975)			
Externalizing behaviors	0.012	0.000	0.024	0.568	0.942	2151
	(0.983)	(1.000)	(0.967)			
Internalizing behaviors	-0.006	0.000	-0.013	0.767	0.948	2151
	(0.992)	(1.000)	(0.985)			

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. Columns (1)–(3) report baseline means for the full sample, control group, and treatment group (with standard deviations in parentheses underneath the means for non-binary variables). Columns (4) and (5) report the p-value and adjusted p-value for the difference in means between control and treatment; adjusted p-values use the Romano and Wolf (2005) procedure with 500 bootstrap samples. Column (6) reports the full sample size.

Table A5: Baseline balance using sample of PMs: households and primary carers characteristics

	Full sample	Control	Treatment	p-value	Adj. p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)
A. Households						
Household size	10.313	10.713	10.032	0.245	0.745	422
	(5.915)	(6.237)	(5.675)			
Number of children 16 or under	4.495	4.816	4.270	0.095	0.441	422
	(3.307)	(3.525)	(3.133)			
Number of children 6 or under	1.597	1.672	1.544	0.453	0.896	422
	(1.724)	(1.767)	(1.695)			
Wealth index (z-score)	-0.064	0.106	-0.183	0.144	0.607	421
	(2.001)	(1.969)	(2.019)			
Main income source:						
Farming own land	0.483	0.494	0.476	0.710	0.976	422
Wage work	0.223	0.201	0.238	0.373	0.882	422
Livestock	0.038	0.040	0.036	0.835	0.976	422
Profits from small enterprise	0.173	0.178	0.169	0.814	0.976	422
B. Primary carers						
Age	36.822	37.402	36.415	0.353	0.890	422
	(10.722)	(11.087)	(10.461)			
Target child's biological mother	0.834	0.828	0.839	0.763	0.990	422
Some education	0.182	0.190	0.177	0.749	0.990	422
Illiterate	0.934	0.920	0.944	0.331	0.890	422
Raven's score	42.160	42.385	42.003	0.813	0.990	422
	(16.304)	(17.782)	(15.217)			
Depression score (1–20)	8.690	8.885	8.552	0.521	0.958	422
	(5.235)	(5.321)	(5.180)			
Self-esteem score	17.725	17.655	17.774	0.734	0.990	422
	(3.539)	(3.799)	(3.352)			

Notes: The sample includes participating mothers' households and primary carers whose children attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. Columns (1)–(3) report baseline means for the full sample, control group, and treatment group (with standard deviations in parentheses underneath the means for non-binary variables). Columns (4) and (5) report the p-value and adjusted p-value for the difference in means between control and treatment; adjusted p-values use the Romano and Wolf (2005) procedure with 500 bootstrap samples. Column (6) reports the full sample size.

Table A6: Baseline balance using sample of PMs: target child characteristics

	Full sample	Control	Treatment	p-value	Adj. p-value	N
	(1)	(2)	(3)	(4)	(5)	(6)
A. Characteristics						
Age (months)	55.803	55.713	55.867	0.872	0.984	422
	(9.642)	(9.410)	(9.819)			
Gender	0.488	0.506	0.476	0.546	0.934	422
Currently in school	0.836	0.833	0.839	0.883	0.984	422
In the last 30 days:						
Any cough	0.557	0.511	0.589	0.117	0.437	422
Any stomach pain	0.528	0.460	0.577	0.018	0.088	422
Any high fever	0.377	0.339	0.403	0.182	0.555	422
Any diarrhea	0.338	0.356	0.325	0.508	0.934	420
B. Cognitive outcomes						
Aggregate score	0.045	0.112	-0.003	0.277	0.625	422
	(1.075)	(1.099)	(1.056)			
Emergent literacy	0.038	0.084	0.005	0.447	0.828	422
	(1.052)	(1.065)	(1.043)			
Emergent numeracy	0.044	0.092	0.010	0.438	0.828	422
	(1.069)	(1.059)	(1.077)			
Executive function	0.019	0.099	-0.037	0.185	0.517	422
	(1.032)	(1.058)	(1.012)			
C. Socioemotional outcomes						
Emotional awareness	0.015	-0.008	0.032	0.705	0.906	422
	(1.067)	(1.073)	(1.064)			
Strengths and Difficulties:						
Emotional and behavioral problems	0.012	0.031	-0.001	0.748	0.906	422
	(0.998)	(1.007)	(0.994)			
Externalising behaviors	0.046	-0.042	0.107	0.133	0.415	422
	(1.004)	(1.017)	(0.993)			
Internalising behaviors	-0.019	0.085	-0.092	0.072	0.269	422
	(0.998)	(1.022)	(0.975)			

Notes: The sample includes participating mothers' children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. Columns (1)–(3) report baseline means for the full sample, control group, and treatment group (with standard deviations in parentheses underneath the means for non-binary variables). Columns (4) and (5) report the p-value and adjusted p-value for the difference in means between control and treatment; adjusted p-values use the Romano and Wolf (2005) procedure with 500 bootstrap samples. Column (6) reports the full sample size.

Table A7: Impacts on parental knowledge, play investments, and parenting behaviors (Mean of Scores)

	I. Knowledge		**II. Play Investments**		**III. Behaviors**	
	Child Development (MoS)	Pre-school Quality (MoS)	Play Materials (MoS)	Any play activities (Binary)	Positive Behaviors (MoS)	Negative Behaviors (MoS)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Treatment among PMs	0.508	0.235	-0.061	0.180	0.236	0.001
	(0.140)	(0.105)	(0.144)	(0.051)	(0.130)	(0.137)
p-val	*0.000*	*0.029*	*0.674*	*0.001*	*0.074*	*0.993*
Adj. p-val	*0.012*	*0.114*	*0.896*	*0.012*	*0.212*	*0.992*
Panel B						
Treatment among non-PMs	-0.051	0.118	-0.192	-0.029	0.050	-0.015
	(0.107)	(0.058)	(0.105)	(0.036)	(0.078)	(0.097)
p-val	*0.639*	*0.045*	*0.072*	*0.431*	*0.526*	*0.875*
Adj. p-val	*0.834*	*0.162*	*0.204*	*0.806*	*0.828*	*0.836*
Panel C						
Difference in treatment	0.559	0.117	0.131	0.209	0.186	0.016
	(0.136)	(0.108)	(0.141)	(0.046)	(0.104)	(0.140)
p-val	*0.000*	*0.279*	*0.356*	*0.000*	*0.077*	*0.907*
Adj. p-val	*0.002*	*0.502*	*0.535*	*0.002*	*0.165*	*0.900*
Control mean endline – non-PMs	0.031	-0.015	-0.091	0.317	0.007	0.011
Control mean endline – PMs	0.079	0.280	-0.060	0.360	0.189	-0.014
Baseline outcome	Yes	No	Yes	Yes	No	No
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151	2151	2,151

Notes: MOS means "Mean of Scores". The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented with standard errors clustered at the community level, and p-values are reported in italics. Throughout we estimate fully interacted models that control for the baseline outcome (if available), strata fixed effects, child's age in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother signed up to participate in the play schemes (during the intervention in the treatment group, in the following year in the control group). At the foot of each column, we report the mean outcome among controls in each subgroup.

Table A8: Parental knowledge about preschool quality

	Teacher discipline	Rote learning	Parents involvement	Building
	(1)	(2)	(3)	(4)
Panel A				
Treatment among PMs	0.028	0.146	-0.038	0.085
	(0.045)	(0.058)	(0.044)	(0.032)
p-val	*0.526*	*0.014*	*0.388*	*0.009*
Adj. p-val	*0.535*	*0.069*	*0.535*	*0.059*
Panel B				
Treatment among non-PMs	0.038	0.085	-0.027	0.015
	(0.028)	(0.027)	(0.029)	(0.015)
p-val	*0.182*	*0.002*	*0.357*	*0.327*
Adj. p-val	*0.413*	*0.014*	*0.493*	*0.493*
Panel C				
Difference in treatment	-0.010	0.061	-0.011	0.070
	(0.039)	(0.056)	(0.047)	(0.034)
p-val	*0.800*	*0.278*	*0.811*	*0.042*
Adj. p-val	*0.949*	*0.465*	*0.949*	*0.152*
Control Mean Endline non-PMs	0.752	0.622	0.817	0.854
Control Mean Endline PMs	0.829	0.737	0.844	0.912
Baseline outcome	No	No	No	No
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented with standard errors clustered at the community level, and p-values are reported in italics. Throughout we estimate fully interacted models that control for the baseline outcome (if available), strata fixed effects, child's age in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother signed up to participate in the play schemes (during the intervention in the treatment group, in the following year in the control group). At the foot of each column, we report the mean outcome among controls in each subgroup.

Table A9: Impacts on play materials and play activities

	A. Play materials															B. Play activities						
	Home toys	Toy pack	Single toy	HH object used as toy	Music toy	Pack of blocks	Single block	Drawing toy pack	Single drawing toy	Toy to help moving	Dolls	Books	Toys to learn shapes	Picture charts	Other	Read	Telling stories	Singing	Going out	Playing	Drawing	Playing with objects
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)
Panel A																						
Treatment among PMs	-0.027	0.014	-0.050	-0.069	-0.030	0.004	0.000	0.015	0.018	-0.008	0.007	-0.003	-0.012	0.017	0.025	0.082	0.101	0.026	0.025	0.027	0.024	0.024
	(0.079)	(0.007)	(0.038)	(0.062)	(0.040)	(0.003)	(0.000)	(0.017)	(0.038)	(0.058)	(0.053)	(0.024)	(0.012)	(0.009)	(0.013)	(0.023)	(0.031)	(0.027)	(0.039)	(0.010)	(0.011)	(0.016)
p-val	*0.735*	*0.056*	*0.197*	*0.270*	*0.454*	*0.284*	*1.000*	*0.382*	*0.630*	*0.888*	*0.891*	*0.896*	*0.342*	*0.064*	*0.067*	*0.001*	*0.002*	*0.340*	*0.514*	*0.009*	*0.031*	*0.141*
Panel B																						
Treatment among non-PMs	0.016	-0.008	0.000	-0.083	-0.034	-0.001	-0.002	-0.018	-0.036	-0.063	-0.032	-0.031	-0.008	-0.003	0.019	-0.009	0.005	-0.024	-0.012	0.004	-0.008	-0.005
	(0.045)	(0.005)	(0.027)	(0.030)	(0.026)	(0.002)	(0.004)	(0.012)	(0.031)	(0.040)	(0.036)	(0.019)	(0.004)	(0.004)	(0.011)	(0.018)	(0.019)	(0.018)	(0.031)	(0.005)	(0.009)	(0.007)
p-val	*0.731*	*0.106*	*0.986*	*0.007*	*0.198*	*0.422*	*0.690*	*0.146*	*0.248*	*0.120*	*0.371*	*0.103*	*0.075*	*0.512*	*0.083*	*0.634*	*0.779*	*0.185*	*0.694*	*0.506*	*0.340*	*0.511*
Panel C																						
Difference in treatment	-0.043	0.022	-0.049	0.014	0.004	0.005	0.002	0.033	0.055	0.054	0.039	0.028	-0.004	0.020	0.006	0.091	0.096	0.050	0.038	0.023	0.032	0.029
	(0.068)	(0.008)	(0.045)	(0.063)	(0.034)	(0.004)	(0.004)	(0.022)	(0.037)	(0.055)	(0.046)	(0.021)	(0.012)	(0.009)	(0.016)	(0.026)	(0.028)	(0.026)	(0.034)	(0.012)	(0.014)	(0.018)
p-val	*0.531*	*0.010*	*0.283*	*0.829*	*0.911*	*0.186*	*0.690*	*0.139*	*0.146*	*0.324*	*0.389*	*0.179*	*0.735*	*0.031*	*0.687*	*0.001*	*0.001*	*0.061*	*0.266*	*0.052*	*0.021*	*0.118*
Control Mean Endline non-PMs	0.291	0.014	0.187	0.641	0.087	0.001	0.008	0.036	0.134	0.287	0.226	0.034	0.009	0.007	0.014	0.086	0.133	0.087	0.102	0.017	0.020	0.014
Control Mean Endline PMs	0.277	0.009	0.194	0.607	0.123	0.002	0.000	0.038	0.121	0.334	0.249	0.031	0.012	0.012	0.014	0.092	0.133	0.097	0.135	0.021	0.024	0.024
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151	2151

Notes: The sample includes children who attended one of the pre-schools randomized into treatment or control at baseline and who were successfully tracked at endline. OLS estimates are presented where standard errors are clustered at the community level, and p-values are reported in italics. Throughout we estimate fully interacted models that control for the baseline outcome (if available), strata fixed effects, child's age measured in months at endline, a gender indicator, and the number of clothing stores in the community, and interactions of all these variables and the treatment indicator with an indicator for whether the child's mother has signed up to participate in the play schemes (during the intervention in the treatment group, in the following year in the control group). At the foot of each column, we report the mean outcome among controls in each subgroup.

Table A10: Impacts of intervention: Controls selected using post-double selection Lasso

	Cognition Aggregate Score	Emotional Awareness Aggregate Score	Behavioral difficulties
	(1)	(2)	(3)
Treatment	0.108**	0.0383	-0.0442
	(0.052)	(0.070)	(0.065)
Control mean endline	0.000	0.000	0.000
Strata Fixed Effects	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes
Observations	2151	2151	2151

Notes: ***, **, and * indicate significance at the 1, 5, and 10 percent level respectively. OLS estimates are reported with standard errors clustered at the village level in parentheses; p-values are reported in italics. Controls are selected using the post-double selection Lasso method (Belloni et al., 2014).

Table A11: Impacts on children's cognitive outcomes (IRT scores)

	Cognition Aggregate Score	Emergent Literacy	Emergent Numeracy	Executive Function
	(1)	(2)	(3)	(4)
Treatment	0.121	0.075	0.094	0.121
	(0.055)	(0.055)	(0.050)	(0.051)
p-val	*0.030*	*0.176*	*0.064*	*0.020*
Adj. p-val	-	*0.096*	*0.064*	*0.032*
Baseline outcome	0.456	0.323	0.439	0.257
	(0.023)	(0.025)	(0.023)	(0.022)
p-val	*0.000*	*0.000*	*0.000*	*0.000*
Control mean endline	0.000	0.000	0.000	0.000
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: All outcomes measured as IRT scores standardised to have mean 0, std 1 in the control group. See Table 3 note for more details.

Table A12: Impacts on children's socioemotional outcomes (IRT scores)

	Emotional Awareness Aggregate Score	Behavioral Difficulties	Internalising Behaviors	Externalising Behaviors
	(1)	(2)	(3)	(4)
Treatment	0.041	-0.096	-0.099	-0.072
	(0.074)	(0.064)	(0.054)	(0.072)
p-val	*0.584*	*0.137*	*0.068*	*0.316*
Adj. p-val	-	-	*0.082*	*0.224*
Baseline outcome	0.233	0.048	0.063	0.044
	(0.024)	(0.026)	(0.022)	(0.029)
p-val	*0.000*	*0.068*	*0.005*	*0.133*
Control mean endline	0.000	0.000	0.000	0.000
Strata Fixed Effects	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151

Notes: All outcomes measured as IRT scores standardised to have mean 0, std 1 in control group. See Table 7 note for more details.

B Measures of child development and parental investments

This appendix section provides a description of the measures of child development and parental investments that we use in the analysis presented in this paper. It establishes the internal validity of these measures and describes how we construct the outcomes we use in the analysis. As part of this, we test the assumptions of scalar and metric invariance across groups of interest.

B.1 Instruments to measure child development and parental investments

Figure B1 lists the instruments used to measure the development of the target child, their younger and older sibling, and parental investments.

Figure B1: Instruments used to measure child development and parental investments

To measure....	We use the following instruments...
Development of target child	IDELA
	Spelke Tasks
	Strengths and Difficulties Questionnaire (SDQ)
Development of younger sibling	CREDI
Development of older sibling	Literacy and numeracy test
	Spelke tasks
Parental investments	Knowledge of Infant Development Inventory (KIDI)
	Knowledge of pre-school quality
	Play materials and play activities (Family Care Indicators)
	Parenting behaviours
Maternal wellbeing	Rosenberg of self-esteem
	SRQ-10 scale of depressive symptoms

When selecting measurement tools to measure the target child's developmental outcomes, we took into account three key considerations. First of all, we selected measures that allowed us to study impacts on specific sub-domains of children's development, rather than overall aggregate measures. For example, we specifically selected instruments that allowed us to separately measure skills related to emergent numeracy, or emergent literacy, as opposed to just overall cognitive development. Second, we wanted to make sure that we used internationally validated instruments that have previously also been used in Ghana and in comparable contexts. A third and final key consideration was to ensure that our child development assessment is complete, so that it covers all important aspects of child development and

also that it was robust to potential “teaching to the test” effects of the Lively Minds curriculum. For these reasons, we used three main instruments: the International Development and Early Learning Assessment (IDELA), tasks developed by Elizabeth Spelke’s Lab ("Spelke Tasks" henceforth), and the Strengths and Difficulties Questionnaire (SDQ).

For the rest of the analysis, we selected, to the best extent possible, already validated measures that had been used in low-income contexts and complemented these measures with newly created measures when available measures did not measure the constructs of interest comprehensively enough.

B.1.1 Instruments to measure the target child’s development

IDELA The IDELA is a free instrument developed by *Save the Children* to measure the early learning and development of children from 3.5 to 6 years old. It was purposefully designed to be internationally applicable and comparable and has already been used and validated in various contexts, including Ghana.

The IDELA assessment takes about 35 minutes per child. It consists of a series of tasks, which are administered and scored by the data collector.

The tasks included in the IDELA cover a range of cognitive and socio-emotional dimensions, which include:

- *emergent numeracy*, i.e. the child’s ability to do basic math which at the age of 4-5 is measured for example by recognition of numbers, shapes and sizes
- *emergent literacy*, i.e. the child’s ability to read or to recognise numbers)
- *fine motor skills*, i.e. skills that use the small muscles, for example picking up small objects or drawing)
- executive functioning, i.e. skills involving mental control and self-regulation, for example ability to memorise, or plan and control impulses, and
- socio-emotional skills, specifically around the child’s expression and management of his or her own emotions, their ability to empathise with others and to establish positive rewarding relationships with them.

The IDELA was extensively piloted before the baseline survey was administered, which led to a few minor changes being made to the original IDELA assessment. These include:

- Item 1(f) was dropped: almost all children got this question wrong during the two pilots
- Item 7: Apples were replaced with mangoes as children in the study area were not familiar with apples
- Item 8: We added a four-piece puzzle to serve as an example for the six-piece puzzle
- Item 11: Picture of a cartoon child crying was changed to an African boy crying
- Item 21: the project adapted a square shape instead of a triangle as most children in this setting were more familiar with a square shape than with a triangle.
- Item 24: The hopping question was replaced with a pencil tap game. 72
- Item 25: this is a health and hygiene question that was added to the usual IDELA items.

Spelke Tasks Based on our third consideration around the comprehensiveness of our child development assessment and the risk of "teaching to the test", we complemented the IDELA tool with a battery of assessments developed for and used in the Ghanaian context by Harvard Cognitive Psychologist Professor Elizabeth Spelke (Spelke lab). With the exception of motor skills, the Spelke tasks measure similar developmental domains as the IDELA.

The tasks that we administered covered a range of skills, including emergent math, emergent literacy, executive function, and emotional awareness. For emergent maths, we administered the following tasks:

- *Geometric Intruder*, which tests whether a child can identify which shape, among four presented to them, presents a unique difference
- *Panamath*, which tests whether the child can quickly determine which of two boxes on a laminated card has more dots (to test numerical estimation)
- *Point to number*, which tests whether the child can identify particular number of toys and, then, to cards with particular arabic numbers on them
- *Extra number* which tests whether the child can identify which of two plates have more toys in them and whether the child can add the number of toys there are on the two plates

For emergent literacy, we administered a single task called *Vocabulary assessment* which tested the size of the child's vocabulary. In this task, the child is presented with a set of laminated cards with

different pictures of objectives. The assessor says a word and asks the child to point to the card with the word on them.

For executive function, we administered the two tasks which involved playing objects on a mat in front of the child:

- *Attention Switching*, which tests whether if the child can find an object after the surveyor taps on the incorrect location.
- *Mental Simulation/Rotation*, which tests whether the child can find objects placed under cups after the cups slide or are switched on the map.

For emotional awareness, we administer a task called *Point to Emotion*, where the child is presented with a number of pictures of human faces showing different emotions and asked to point to the image showing the emotions described by the assessor.

Strengths and Difficulties Questionnaire The third instrument we administered to measure child developmental outcomes is the Strengths and Difficulties Questionnaire or SDQ (Goodman, 2001). We used the parent-report version of the SDQ, suitable for children aged 4-17 years. The SDQ comprises of 25 items, each describing a particular child behaviour. Possible responses are on a 3-point Likert scale: (0) not true, (1) somewhat true, and (2) certainly true. Those 25 items are made of 5 items to describe each of five domains: conduct problems, hyperactivity, emotional problems, peer problems, and pro-sociality. The conduct problems and hyperactivity scales together make a measure of externalising problems, while the emotional problems and peer problems scales make a measure of internalising problems. Our analysis focused on 20 items which relate to items which relate to internalizing and externalizing problem behaviors.

The SDQ (Goodman, 1997) was initially developed and validated in Western HIC settings (Goodman, 2001; Henrich et al., 2010), but has been used internationally with versions available in over 90 languages (Youth in Mind, 2020), including in Sub-Saharan Africa and Ghana (e.g., Brown et al., 2025; Hoosen et al., 2018).

B.1.2 Instrument to measure the younger sibling's development

We measured the development of younger siblings using the Caregiver-Reported Early Development Index (CREDI) short form (McCoy et al., 2017). The CREDI are a set of population-level measures

designed to assess early childhood development for children from birth to age three using caregiver reports. Developed to be feasible for use in large-scale surveys and in low-resource settings, the CREDI focus on core developmental domains including cognition, language, motor skills, and socio-emotional development. The instruments are administered through brief, age-specific questionnaires that ask caregivers about observable behaviors and milestones, rather than requiring direct child assessment. The CREDI Short Form consists of 20 caregiver-reported items for each child, designed as a fixed-length set to rapidly assess overall early childhood developmental status in population-level surveys and monitoring efforts. This contrasts with the Long Form, which includes up to 100 items and allows more detailed, domain-specific scoring

B.1.3 Instruments used to measure older siblings' development

We measured literacy and numeracy of older siblings with an adapted version of a test previously used in the Ghanaian context and developed by the Ministry of Education officials from the National Council for Curriculum and Assessments (NaCCA). The test items are aligned with the Ghana Education Service curriculum. The instrument includes tasks related to subtraction, addition and fractions.

At endline we complemented these literacy and numeracy tests with a set of tasks developed in the Spelke lab. These tasks are aimed to assess math (*Geometric intruder*), literacy (*Vocabulary assessment*), emotional awareness (*Point to emotion*), and executive function. For executive function, the task tested the child's ability to hit the key on the same side or opposite side as an object.

B.1.4 Instruments to measure parental knowledge

Knowledge of Infant Development Inventory (KIDI) We measured the primary carer's knowledge of stimulation and care practice and her beliefs regarding the importance of these for children's development using the Knowledge of Infant Development (KIDI).The Knowledge of Infant Development Inventory (KIDI) is a widely used self-report measure of parental knowledge about child developmental processes, norms and milestones, caregiving practices, and health and safety guidelines. In its classic form, the KIDI consists of 75 items, covering a broad range of content relevant to infants and young children, with respondents indicating whether statements are agree, disagree, or unsure (and in some milestone items, options about the age at which behaviors are expected). In the study, we used an abbreviated six-item of the KIDI.

Knowledge about pre-school quality The second measure of parental knowledge hones in specifically on parental knowledge relating to pre-school quality including most productive care and teaching strategies, key features of pre-school infrastructure, and parental involvement. To assess this, we presented 4 pairs of vignettes designed by a local artist and, for each pair, asked parents to point to the one they thought was associated with higher-equality education.

B.1.5 Instruments to measure play materials and play activities

We used the Family Care Indicators (FCI) instrument which was derived from the Home Observations for Measurement of the Environment for use in large-scale population surveys (Hamadani et al.,2010). The scale captures the availability of play and learning materials in the household, as well as educational and play activities conducted with the child in the previous 3 days by any household member over the age of 15. Enumerators recorded whether each of 13 categories of play and learning materials were available. Maternal report was used to establish whether and how many times each of the several learning and play activities took place over the preceding 3 days.

B.1.6 Instrument to measure positive and negative parenting behaviours

Parenting behaviours were assessed using two observational approaches conducted by trained enumerators. First, enumerators directly observed mothers' teaching behaviors while they interacted with their child during two structured activities. Mothers were instructed on how to carry out two simple guessing and sorting games and then asked to play these games with their child, providing an opportunity for the enumerator to assess teaching style. Enumerators recorded the presence or absence of 15 specific behaviors reflecting instructional practices during the activities (coded 0 = not observed; 1 = observed). These behaviors captured both supportive and less supportive teaching strategies, such as adapting instructional approaches when initial strategies were ineffective, tailoring guidance to the child's prior actions, limiting the child's opportunities for independent problem-solving, or displaying low confidence while completing the task.

Second, enumerators documented mother–child interactions observed throughout the interview process. During the survey, enumerators coded whether mothers engaged in nonviolent or violent disciplinary practices toward the child (0 = no; 1 = yes). Nonviolent discipline included behaviors such as removing privileges, explaining inappropriate behavior, or redirecting the child's attention. Violent

discipline encompassed psychological aggression (e.g., shouting or using disparaging language) and physical punishment (e.g., shaking, spanking, hitting, or slapping the child, with or without an object). These items were drawn from a subset of the discipline measures used in the UNICEF Multiple Indicator Cluster Surveys (MICS) (Bornstein et al., 2023), and enumerators were trained to systematically observe and code these behaviors during the course of the survey.

We grouped items from both sets of data about parenting behaviours into two groups: one group of positive parenting behaviours and approaches and one group of negative parenting behaviours and approaches.

B.1.7 Instruments to measure maternal well-being

Mothers' well-being was measured in two ways. First, the Rosenberg Self-Esteem scale was administered (Rosenberg, 1965). This 10-item scale is globally one of the most widely used measures of self-esteem including in LMICs (Schmitt & Allik, 2005). Responses are on a 4-point Likert scale 1= strongly disagree and 4= strongly agree.

The second measure of mothers' wellbeing captured mental health using the SRQ-20. psychological distress scale. This scale combines brevity with strong psychometric properties and has been used globally as part of the WHO World Mental Health Surveys (Kessler, 2002). It consists of 10 items asking about the frequency with which respondents experience a range of symptoms. Responses are on a 5-point Likert scale 1= none of the time and 5= all of the time.

B.2 Constructing measures of child development and parental investments

Constructing measures of child development and parental investments we use in the analysis involved two main steps. First, we mapped individual items to constructs of interest and validated such mapping by checking the construct's internal consistency. Second, we created aggregate scores for each of the constructs.

B.2.1 Internal consistency of mapping between measures and constructs

Figure B2 shows how we mapped each of the items from the IDELA, set of Spelke Tasks, and SDQ to create three main constructs of child development: cognitive skill, emotional awareness, and emotional and behavioural problems. This mapping is entirely based on the three tests' developers' guidance about

what underlying skill or construct each item measures.

As is clear in the table, items both from the IDELA and from the set of Spelke Tasks measure similar cognitive sub-constructs, including emergent literacy, emergent numeracy, and executive function, as well as emotional awareness. Emotional and behavioural problems are uniquely measured using SDQ items measuring conduct problems, emotional problems, hyperactivity, and peer problems. The items measuring emotional and behavioural difficulties map onto two sub-constructs, one measuring externalising behaviours (SDQ items measuring conduct problems and hyperactivity) and one measuring internalising behaviours (SDQ items measuring peer problems and emotional problems).

While this mapping follows the test developers' own guidelines about what dimension of child development each item aims to measure, we also empirically validate the mapping by examining the correlations between the items assigned to the same (sub-) construct and calculating their Cronbach's alpha, a measure of internal consistency which assess how well the group of measures actually measure a similar, unique construct.

Figure B3 shows the correlation matrix between all the items included when measuring cognitive skill. Their Cronbach alpha is .879, pointing towards very strong internal consistency. Figure B4, Figure B5, and Figure B6 report the correlation matrices of the subsets of elements used to measure the subdimensions of cognitive skill, namely emergent literacy, emergent numeracy, and executive function. Their internal validity as measured by the Cronbach's alpha is lower than for the overall cognitive skill, and especially so for executive function. While we still report results on each of the cognitive sub-constructs, the fact that internal consistency is lower for the cognitive sub-constructs than for the overall cognitive construct motivates us to place further confidence on the results for the overall cognitive construct.

We perform a similar exercise for the two socio-emotional dimensions we consider: emotional awareness (Figure B7) and emotional and behavioural problems (Figure B8) . The Cronbach's alpha for each of these constructs is 0.66 and 0.75, which are deemed reasonable in the psychometrics literature. The last two figures (Figure B9 and Figure B10) report the same analysis for the sub-dimensions of emotional and behavioural problems. The internal validity of the construct measuring externalising behaviours is the lowest (α=.57) and slightly higher for internalising behaviours (α=.62).

We then turn to examining the correlation between items measures each of the constructs measuring parental knowledge, play investments and parenting behaviours in Figure B11 to **??**. Internal consistency

is reasonable or strong for parental knowledge of child development (measured by the KIDI), positive and negative parenting behaviours. It is lower for pre-school quality ($\alpha = 0.32$) and play materials ($\alpha = .45$) measured by the FCI scale.

Finally, the items measuring depressive symptoms from the SRQ-20 offer very high internal consistency ($\alpha = .88$). The items measuring self-esteem from the Rosenberg scale offer lower, but still reasonable internal consistency ($\alpha = .60$).

B.2.2 Constructing aggregate scores for constructs of interest

The analysis presented above suggests that, with a few exceptions, the items we designate to measure the key constructs in our analysis have relatively good internal consistency and can therefore be grouped to measure a single, underlying construct. We now discuss the approaches we can and do take to construct an aggregate score measuring that particular construct from the multiple items measuring it.

To fix ideas, it is useful to explicitly model the relationship between the underlying construct, say θ, and the observed measures of it which contain information about the construct as well as some measurement error. Assume for all $j = 1, \ldots, J$, we can write $Y_{i,j}$ the j-th measure of the underlying θ for individual i as a linear function of the underlying construct and measurement error $\epsilon_{i,j}$:

$$Y_{i,j} = \alpha_j + \lambda_j \theta_i + \epsilon_{i,j} \qquad \text{(Index 1)}$$

where α_j is referred to as an intercept and λ_j is referred to as a factor loading. The model above is a simple latent factor model.

The most straightforward approach - and that recommended by the developers of the test we use in the trial (IDELA, Spelke Tasks, SDQ) - to aggregating θ_i is to average across the J measures of such construct. For IDELA tasks, for example, this amounts to a)for each task, calculating the percentage of sub-tasks the child gets correct across all sub-tasks pertaining to that tasks and b) averaging those scores across all tasks pertaining to one construct. In the framework above, this comes down to imposing that the factor loadings on the construct θ are all equal to each other for all measures $j = 1, dots, J$.

Another approach, which is very common in the psychometrics literature and is increasingly common in the economics literature, is to estimate the parameters of the model above so as to best fit the moments of the joint distribution of Y_j's. The model described in *Equation Index*1 is identified under a number of standard conditions and normalisations, including that $J3$, a location normalisation for exam-

ple via the assumption that $E(\theta) = 0$, a scale normalisation for example via the assumption that $\lambda_1 = 1$, and an assumption that the measurement error is i.i.d and uncorrelated across measures. Under these assumptions, the model is identified and can be estimated via standard estimation techniques, such as generalised least squares.

Our paper is primarily concerned with identifying gaps in such constructs - we aim to measure gaps in child development or parental investments between treated and control units, but also between Participating Mothers and non-Participating Mothers. In order to do so validly, it is critical that *measurement invariance* holds, that is that the observed tests or scales we use to measure underlying constructs actually do measure those constructs in the same across the different groups we aim to compare. If measurement invariance does not hold, observed score differences may partly reflect changes in the measurement process itself (for example, differential item functioning induced by the intervention), rather than true differences in skills between the groups - and meaningful comparisons between groups cannot hold.

Measurement invariance is commonly discussed in terms of increasingly restrictive levels. **Metric invariance** (sometimes called weak invariance) requires that factor loadings are equal across groups. This means that each assessment item contributes to the underlying skill in the same proportional way for treated and control children. When metric invariance holds, differences in relationships—such as regression coefficients linking skills to later outcomes—can be meaningfully compared across groups, because the scale of the latent skill is the same. **Scalar invariance** (or strong invariance) adds the requirement that item intercepts are equal across groups. This ensures that, for a given level of the underlying skill, children in different groups have the same expected item scores. Scalar invariance is crucial for comparing average levels of latent skills across treated and control groups, since it rules out systematic shifts in item difficulty that differ by treatment status. **Strict invariance** is the strongest form and additionally requires equality of item residual variances, implying that the amount of measurement error is the same across groups. While strict invariance is not necessary for the type of conclusions we would want to draw in this paper, it supports the strongest claims of full comparability of observed scores and latent variables.

In practice, full invariance at all levels often fails, especially in applied settings such as early childhood assessments, where interventions may differentially affect specific items (e.g., familiarity with testing formats or task-specific skills). This motivates the concept of **partial invariance**, where invariance is imposed on a subset of items while allowing others to vary across groups. Partial scalar or

metric invariance can still permit meaningful comparisons of latent means or relationships, provided that enough items remain invariant to anchor the scale of the latent construct. In an experimental evaluation, evidence of partial invariance can be informative in its own right, as non-invariant items may reveal mechanisms through which the treatment affects performance on specific tasks, even if the underlying skill is unchanged.

Testing for measurement invariance is typically carried out as a sequence of increasingly restrictive model comparisons. The starting point is a **configural (fully unconstrained) model**, estimated separately but simultaneously across treated and control groups, in which the same factor structure is imposed but all key parameters—factor loadings, item intercepts, and residual variances—are allowed to differ by group. This model establishes a baseline: it assesses whether the overall measurement framework is plausible in each group and provides a reference level of model fit against which more restrictive specifications can be evaluated.

From this baseline, invariance is tested by imposing equality constraints step by step and examining whether these restrictions lead to a statistically and substantively meaningful deterioration in model fit. The first step is to test **metric invariance** by constraining factor loadings to be equal across groups. This restriction implies that items relate to the underlying skill in the same way for treated and control children. The fit of the metric model is then compared to the configural model, typically using likelihood ratio tests or equivalent fit statistics. If the fit does not worsen significantly, metric invariance is supported.

Next, **scalar invariance** is tested by additionally constraining item intercepts to be equal across groups. This step evaluates whether, conditional on the latent skill, children in different groups have the same expected item scores. Again, the scalar model is compared to the metric model to assess whether the added restrictions are empirically defensible. Finally, *strict invariance* is tested by also constraining residual variances to be equal across groups, implying equal measurement error across treatment status.

When full invariance at a given stage is rejected, the analysis proceeds by searching for *partial invariance* rather than abandoning cross-group comparability altogether. In practice, this involves selectively relaxing equality constraints on a subset of parameters while maintaining constraints on others. To guide this process, we examine *score tests* (or equivalently, modification indices), which quantify the expected improvement in model fit—often expressed as the change in the log-likelihood—associated with freeing a particular constrained parameter. Parameters with the largest scores indicate the constraints that

are most strongly violated by the data. We then free these parameters sequentially, beginning with those associated with the greatest score, and re-estimate the model until an acceptable balance between model fit and parsimony is achieved. This procedure allows the latent skill to remain anchored by invariant items while accommodating targeted departures from invariance that may arise from treatment-induced changes in specific assessment tasks.

In our specific context, we need to test for measurement invariance across four groups: Participating mothers in the treated and control communities, and non-Participating Mothers in the treated and control communities. Note that, if anything, we would expect more deviations from measurement invariance for the measures that are self-reported rather than those assessed by the surveyor.

We describe the results of our tests for measurement invariance for each construct of interest in Figure B18 to Figure B21. Overall, there is evidence of some measurement invariance for emergent numeracy, the overall cognitive construct (which includes emergent numeracy), externalising behaviours, behavioural problems (which include externalising behaviours), positive and negative parenting, primary carer's depressive symptoms and self-esteem. However, for all those constructs, it is possible to find a model where partial metric invariance holds, that is where the fit of the model freeing up a few factor loadings and/or intercepts is as good as the fit of the fully unconstrained model.

For constructs where partial rather than strong measurement invariance holds, we estimate the impacts of Lively Minds among the PM and non-PM groups using two different ways of aggregating the measures. The first, most simple one, is to take an average of raw scores measuring each construct. This approach essentially assumes equal factor loadings across all measures and across all groups. The second is to estimate factor scores for the underlying constructs based on the factor model where partial invariance holds. Overall, as as shown in outcomes instead of our favorite measures. As is clear from Figure B22, the results are extremely similar between the two approaches. For the sake of simplicity and transparency, we present, in the main text, impacts based on averages of raw scores.

Figure B2: Instruments used to measure the target child's development

Constructs	Sub-constructs	Measurements
Cognitive skill	**Emergent math**	Geometric intruder (Spelke task) Comparison by size and length (Idela 2) Sorting and classification (Idela 3) Shape identification (Idela 4) Puzzle completion (Idela 8) Panamath (Spelke task) Point to number (Spelke task) Number identification (Idela 5) One to one correspondence (Idela 6) Extra number (Spelke task) Addition and subtraction (Idela 7)
	Emergent literacy	Vocabulary assessment (Spelke task) Expressive vocabulary (Idela 15) Print awareness (Idela 16) Letter identification (Idela 17) First letter sounds (Idela 18) Emergent writing (Idela 19) Oral comprehension (Idela 20)
	Executive function	Attention switching (Spelke task) Mental simulation / rotation (Spelke task) Short term memory (Idela 13) Inhibitory control (Idela 14)
Emotional awareness		Point to emotion (Spelke task) Self-awareness (Idela 1) Friends (Idela 9) Emotional awareness/regulation (Idela 10) Empathy/perspective taking (Idela 11) Solving conflict (Idela 12)
Emotional and behavioural problems	**Externalising behaviours**	SDQ tempers, obeys, fights, lies, steals SDQ restless, fidgety, distracted, thinks before acting, good attention
	Internalising behaviours	SDQ somatic symptoms, worries, unhappy, SDQ solitary, has good friend, generally liked, bullied, help out

	IDELA 6	IDELA 2	IDELA 3	IDELA 4	IDELA 5	IDELA 7	IDELA 8	Point to number	Panamath	Extra number	Geometric intruder	IDELA 15	IDELA 16	IDELA 17	IDELA 18	IDELA 19	IDELA 20	Vocab assess-ment	IDELA 14	IDELA 13	Mental Simu-lation	Attention switching
IDELA 6	1.000																					
IDELA 2	0.228	1.000																				
IDELA 3	0.285	0.207	1.000																			
IDELA 4	0.327	0.209	0.321	1.000																		
IDELA 5	0.582	0.179	0.236	0.312	1.000																	
IDELA 7	0.550	0.280	0.308	0.293	0.491	1.000																
IDELA 8	0.427	0.228	0.327	0.347	0.368	0.398	1.000															
Point to number	0.552	0.228	0.272	0.349	0.623	0.481	0.398	1.000														
Panamath	0.244	0.175	0.156	0.181	0.177	0.244	0.202	0.256	1.000													
Extra number	0.307	0.171	0.171	0.194	0.279	0.284	0.241	0.332	0.192	1.000												
Geometric intruder	0.419	0.254	0.264	0.346	0.346	0.406	0.362	0.482	0.299	0.299	1.000											
IDELA 15	0.313	0.217	0.257	0.393	0.298	0.371	0.338	0.348	0.201	0.167	0.341	1.000										
IDELA 16	0.342	0.149	0.274	0.301	0.339	0.333	0.336	0.327	0.257	0.201	0.297	0.339	1.000									
IDELA 17	0.469	0.151	0.172	0.223	0.767	0.385	0.303	0.484	0.124	0.213	0.251	0.208	0.281	1.000								
IDELA 18	0.226	0.100	0.160	0.240	0.243	0.212	0.249	0.205	0.108	0.146	0.159	0.240	0.204	0.229	1.000							
IDELA 19	0.436	0.139	0.226	0.347	0.491	0.372	0.382	0.452	0.173	0.221	0.333	0.398	0.343	0.439	0.248	1.000						
IDELA 20	0.292	0.243	0.206	0.110	0.237	0.343	0.244	0.296	0.154	0.190	0.293	0.277	0.196	0.173	0.079	0.145	1.000					
Vocabulary assessm	0.390	0.326	0.274	0.296	0.365	0.401	0.348	0.478	0.327	0.335	0.470	0.344	0.281	0.292	0.187	0.320	0.368	1.000				
IDELA 14	0.429	0.255	0.302	0.289	0.321	0.388	0.355	0.400	0.240	0.216	0.400	0.361	0.293	0.250	0.177	0.292	0.297	0.386	1.000			
IDELA 13	0.289	0.215	0.204	0.271	0.251	0.305	0.274	0.300	0.171	0.228	0.277	0.345	0.260	0.199	0.186	0.279	0.211	0.283	0.344	1.000		
Mental Simulation	0.107	0.113	0.039	0.016	0.073	0.102	0.083	0.082	0.025	0.122	0.073	0.050	0.013	0.052	0.037	0.021	0.058	0.127	0.081	0.101	1.000	
Attention switching	0.111	0.086	0.092	0.089	0.120	0.134	0.112	0.142	0.069	0.103	0.143	0.121	0.049	0.098	0.102	0.072	0.084	0.202	0.142	0.071	0.195	1.000
Cronbach alpha	**0.879**																					

Figure B3: Correlation between items used to measure cognitive skill

	IDELA 15	IDELA 16	IDELA 17	IDELA 18	IDELA 19	IDELA 20	Vocabulary assessment
IDELA 15	1.000						
IDELA 16	0.339	1.000					
IDELA 17	0.208	0.281	1.000				
IDELA 18	0.240	0.204	0.229	1.000			
IDELA 19	0.398	0.343	0.439	0.248	1.000		
IDELA 20	0.277	0.196	0.173	0.079	0.145	1.000	
Vocabulary assessment	0.344	0.281	0.292	0.187	0.320	0.368	1.000
Cronbach alpha:	**0.679**						

Figure B4: Correlation between items used to measure emergent literacy

	IDELA 6	IDELA 2	IDELA 3	IDELA 4	IDELA 5	IDELA 7	IDELA 8	Point to number	Panamath	Extra number	Geometric intruder
IDELA 6	1.000										
IDELA 2	0.228	1.000									
IDELA 3	0.285	0.207	1.000								
IDELA 4	0.327	0.209	0.321	1.000							
IDELA 5	0.582	0.179	0.236	0.312	1.000						
IDELA 7	0.550	0.280	0.308	0.293	0.491	1.000					
IDELA 8	0.427	0.228	0.327	0.347	0.368	0.398	1.000				
Point to number	0.552	0.228	0.272	0.349	0.623	0.481	0.398	1.000			
Panamath	0.244	0.175	0.156	0.181	0.177	0.244	0.202	0.256	1.000		
Extra number	0.307	0.171	0.171	0.194	0.279	0.284	0.241	0.332	0.192	1.000	
Geometric intruder	0.419	0.254	0.264	0.346	0.346	0.406	0.362	0.482	0.299	0.299	1.000
Cronbach alpha:	**0.833**										

Figure B5: Correlation between items used to measure emergent numeracy

	IDELA 14	IDELA 13	Mental Simulation	Attention switching
IDELA 14	1.000			
IDELA 13	0.344	1.000		
Mental Simulation	0.081	0.101	1.000	
Attention switching	0.142	0.071	0.195	1.000
Cronbach alpha	**0.417**			

Figure B6: Correlation between items used to measure executive function

	IDELA 9	IDELA 1	IDELA 10	IDELA 11	IDELA 12	Point to emotion
IDELA 9	1.000					
IDELA 1	0.254	1.000				
IDELA 10	0.233	0.257	1.000			
IDELA 11	0.233	0.271	0.359	1.000		
IDELA 12	0.192	0.291	0.313	0.325	1.000	
Point to emotion	0.127	0.197	0.255	0.225	0.169	1.000
Cronbach alpha	**0.656**					

Figure B7: Correlation between items used to measure emotional awareness

	SDQ Restless	SDQ Fidgety	SDQ Distracted	SDQ Thinks before acting	SDQ Good attention	SDQ Tempers	SDQ Fights or bullies	SDQ Lies or cheats	SDQ Steals	SDQ Obeys	SDQ Many fears	SDQ Somatic symptoms	SDQ Worries	SDQ Unhappy	SDQ Nervous in new situations	SDQ Solitary	SDQ Picked on or bulled	SDQ Better with adults	SDQ Has good friend	SDQ Generally liked
SDQ Restless	1.000																			
SDQ Fidgety	0.236	1.000																		
SDQ Distracted	0.269	0.300	1.000																	
SDQ Thinks before acting	-0.071	-0.101	-0.115	1.000																
SDQ Good attention	-0.027	0.029	0.006	0.358	1.000															
SDQ Tempers	0.243	0.220	0.198	-0.043	0.031	1.000														
SDQ Fights or bullies	0.233	0.206	0.192	-0.096	-0.052	0.231	1.000													
SDQ Lies or cheats	0.144	0.178	0.144	0.002	0.013	0.164	0.195	1.000												
SDQ Steals	0.019	0.172	0.101	0.018	0.075	0.094	0.077	0.288	1.000											
SDQ Obeys	-0.039	-0.055	-0.088	0.253	0.258	-0.016	0.014	0.093	0.198	1.000										
SDQ Many fears	0.163	0.259	0.300	-0.165	-0.191	0.143	0.137	0.162	0.055	-0.119	1.000									
SDQ Somatic symptoms	0.207	0.174	0.209	-0.028	-0.029	0.153	0.158	0.157	0.052	-0.039	0.224	1.000								
SDQ Worries	0.155	0.320	0.217	-0.023	0.005	0.172	0.186	0.206	0.155	0.056	0.258	0.172	1.000							
SDQ Unhappy	0.130	0.204	0.241	-0.053	-0.048	0.166	0.215	0.194	0.153	0.054	0.222	0.182	0.270	1.000						
SDQ Nervous in new situations	0.231	0.317	0.383	-0.112	-0.037	0.153	0.206	0.159	0.126	-0.031	0.289	0.233	0.256	0.232	1.000					
SDQ Solitary	0.080	0.116	0.129	-0.072	-0.109	0.111	0.127	0.136	0.098	-0.063	0.185	0.140	0.228	0.195	0.176	1.000				
SDQ Picked on or bulled	0.167	0.136	0.202	-0.125	-0.071	0.190	0.290	0.224	0.110	-0.068	0.188	0.146	0.130	0.229	0.175	0.096	1.000			
SDQ Better with adults	0.016	0.094	0.064	-0.109	-0.153	-0.015	0.069	0.077	0.111	-0.047	0.210	0.083	0.135	0.095	0.099	0.158	0.017	1.000		
SDQ Has good friend	-0.143	-0.057	-0.062	0.159	0.183	-0.056	-0.084	-0.012	0.130	0.225	-0.146	-0.063	-0.018	-0.045	-0.032	-0.113	-0.045	-0.034	1.000	
SDQ Generally liked	-0.152	-0.119	-0.166	0.193	0.184	-0.080	-0.038	0.041	0.169	0.323	-0.175	-0.113	0.016	-0.010	-0.107	-0.038	-0.065	0.033	0.269	1.000
Cronbach alpha	**0.748**																			

Figure B8: Correlation between items used to measure emotional and behavioural problems

	SDQ Restless	SDQ Fidgety	SDQ Distracted	SDQ Thinks before acting	SDQ Good attention	SDQ Tempers	SDQ Fights or bullies	SDQ Lies or cheats	SDQ Steals	SDQ Obeys
SDQ Restless	1.000									
SDQ Fidgety	0.236	1.000								
SDQ Distracted	0.269	0.300	1.000							
SDQ Thinks before acting	-0.071	-0.101	-0.115	1.000						
SDQ Good attention	-0.027	0.029	0.006	0.358	1.000					
SDQ Tempers	0.243	0.220	0.198	-0.043	0.031	1.000				
SDQ Fights or bullies	0.233	0.206	0.192	-0.096	-0.052	0.231	1.000			
SDQ Lies or cheats	0.144	0.178	0.144	0.002	0.013	0.164	0.195	1.000		
SDQ Steals	0.019	0.172	0.101	0.018	0.075	0.094	0.077	0.288	1.000	
SDQ Obeys	-0.039	-0.055	-0.088	0.253	0.258	-0.016	0.014	0.093	0.198	1.000
Cronbach alpha	**0.570**									

Figure B9: Correlation between items used to measure externalising behaviours

	SDQ Many fears	SDQ Somatic symptoms	SDQ Worries	SDQ Unhappy	SDQ Nervous in new situations	SDQ Solitary	SDQ Picked on or bulled	SDQ Better with adults	SDQ Has good friend	SDQ Generally liked
SDQ Many fears	1.000									
SDQ Somatic symptoms	0.224	1.000								
SDQ Worries	0.258	0.172	1.000							
SDQ Unhappy	0.222	0.182	0.270	1.000						
SDQ Nervous in new situ	0.289	0.233	0.256	0.232	1.000					
SDQ Solitary	0.185	0.140	0.228	0.195	0.176	1.000				
SDQ Picked on or bulled	0.188	0.146	0.130	0.229	0.175	0.096	1.000			
SDQ Better with adults	0.210	0.083	0.135	0.095	0.099	0.158	0.017	1.000		
SDQ Has good friend	-0.146	-0.063	-0.018	-0.045	-0.032	-0.113	-0.045	-0.034	1.000	
SDQ Generally liked	-0.175	-0.113	0.016	-0.010	-0.107	-0.038	-0.065	0.033	0.269	1.000
Cronbach alpha	**0.621**									

Figure B10: Correlation between items used to measure internalising behaviours

	Parent role in learning	Read/write important	Parent can support	Learning through play	Parent can engage while work	Praising important
Parent role in learning	1.000					
Read/write important	0.379	1.000				
Parent can support	0.272	0.325	1.000			
Learning through play	0.248	0.265	0.317	1.000		
Parent can engage while work	0.175	0.218	0.290	0.344	1.000	
Praising important	0.250	0.294	0.314	0.287	0.252	1.000
Cronbach alpha	**0.688**					

Figure B11: Correlation between items used to measure knowledge of child development (KIDI meausures)

	Discipline	Building	Rote base learning	Parents involved
Discipline	1.000			
Building	0.138	1.000		
Rote base learning	0.193	0.172	1.000	
Parents involved	0.032	-0.001	0.087	1.000
Cronbach alpha	**0.325**			

Figure B12: Correlation between items used to measure pre-school quality

	Home toys	Toy pack	Single toy	Toy HH object	Music toy	Block packs	Single blocks	Drawing pack	Single drawing set	Toy to move	Dolls	Books	Toys to learn shapes	Picture books	Other
Home toys	1.000														
Toy pack	0.072	1.000													
Single toy	0.095	-0.024	1.000												
Toy HH object	0.056	0.002	0.028	1.000											
Music toy	0.145	0.005	0.113	0.143	1.000										
Block packs	0.031	0.106	-0.018	0.028	0.073	1.000									
Single blocks	-0.013	0.042	-0.009	0.038	0.033	-0.003	1.000								
Drawing pack	0.030	0.022	0.047	0.049	0.082	-0.007	0.015	1.000							
Single drawing set	0.125	-0.008	0.031	0.084	0.081	-0.015	0.003	0.138	1.000						
Toy to move	0.119	-0.012	0.063	0.150	0.230	0.030	-0.027	0.010	0.178	1.000					
Dolls	-0.090	-0.024	0.031	0.074	0.130	0.039	-0.031	0.036	0.130	0.127	1.000				
Books	0.184	0.047	0.062	0.093	0.311	0.132	-0.015	0.061	0.165	0.213	0.101	1.000			
Toys to learn shapes	-0.001	0.030	0.013	0.026	0.033	-0.004	0.051	0.057	0.073	0.029	0.036	0.139	1.000		
Picture books	0.047	0.082	0.011	0.013	0.079	-0.003	0.058	0.011	0.074	0.023	-0.011	0.159	0.151	1.000	
Other	-0.041	0.021	0.034	-0.050	-0.025	-0.004	-0.010	-0.002	-0.035	-0.034	0.001	0.000	-0.012	0.034	1.000
Cronbach alpha	**0.449**														

Figure B13: Correlation between items used to measure play materials

		Took away privileges	Shook	Shouted	Gave smth else to do	Spanked, hit	Hit on bottom	Called dumb	Slapped	Displayed anger	Disinterested	Acted intrusively	Took away privilege	Explained why wrong	Shook	Shouted	Gave smth else to do	Spanked, hit	Hit on bottom	Called dumb
During task,	**Took away privileges**	1.000																		
	Shook	-0.026	1.000																	
	Shouted	0.111	0.083	1.000																
	Gave smth else to do	0.142	0.132	0.166	1.000															
	Spanked, hit	0.030	0.110	0.270	0.055	1.000														
	Hit on bottom	0.078	0.133	0.193	0.115	0.283	1.000													
	Called dumb	0.027	0.121	0.270	0.048	0.276	0.194	1.000												
	Slapped	0.038	0.120	0.173	0.075	0.276	0.306	0.207	1.000											
	Displayed anger	0.051	0.014	0.295	0.016	0.265	0.150	0.221	0.165	1.000										
	Disinterested	0.018	0.012	0.032	0.004	0.086	0.042	0.056	0.038	0.141	1.000									
	Acted intrusively	0.120	0.041	0.136	0.011	0.111	0.017	0.148	0.054	0.142	0.083	1.000								
During interview,	**Took away privilege**	0.742	-0.041	0.101	0.132	0.053	0.049	0.026	0.045	0.046	-0.003	0.121	1.000							
	Explained why wrong	0.205	0.114	0.098	0.170	-0.001	0.021	0.039	0.032	-0.036	-0.036	0.004	0.268	1.000						
	Shook	-0.052	0.492	0.121	0.110	0.046	0.047	0.075	0.098	0.050	-0.016	-0.010	-0.021	0.177	1.000					
	Shoued	0.060	0.053	0.486	0.090	0.177	0.084	0.151	0.099	0.186	-0.004	0.147	0.168	0.135	0.113	1.000				
	Gave smth else to do	0.125	0.140	0.129	0.445	0.070	0.057	0.068	0.060	0.024	0.001	0.077	0.168	0.212	0.211	0.253	1.000			
	Spanked, hit	0.010	0.044	0.198	0.021	0.388	0.105	0.124	0.172	0.169	0.048	0.076	0.068	0.002	0.112	0.313	0.164	1.000		
	Hit on bottom	0.008	0.028	0.127	-0.003	0.139	0.280	0.059	0.062	0.099	0.079	0.034	0.078	-0.013	0.141	0.172	0.065	0.308	1.000	
	Called dumb	-0.014	0.093	0.238	0.023	0.195	0.113	0.529	0.108	0.184	0.048	0.150	0.041	0.046	0.084	0.270	0.043	0.223	0.125	1.000
	Cronbach alpha	**0.693**																		

Figure B14: Correlation between items used to measure positive parenting behaviours

		Praised	Showed signs of affection	Rewarded	Here to provide help	Responded positiviely	Focused the child	Provided instruc-tions	Tried new strategies	Provided feedback	Did not give up	Shy, under confident in task	Set rules/expec tations	Broke task in small steps	Turn task into game	Used descriptive words	Told child name of things/ people	Praised	Conveyed positive feelings	Caressed/ kissed	Kept child in visual range	Responded verball to vocali-sations
During task,	Praised	1.000																				
	Showed signs of affection	0.265	1.000																			
	Rewarded	0.101	0.087	1.000																		
	Here to provide help	0.181	0.170	0.093	1.000																	
	Responded positiviely	0.340	0.130	0.040	0.180	1.000																
	Focused the child	0.217	0.106	0.015	0.149	0.428	1.000															
	Provided instructions	0.139	0.140	0.063	0.155	0.213	0.187	1.000														
	Tried new strategies	0.206	0.172	0.070	0.244	0.246	0.254	0.200	1.000													
	Provided feedback	0.342	0.154	0.028	0.191	0.377	0.310	0.247	0.290	1.000												
	Did not give up	0.188	0.069	0.037	0.126	0.326	0.457	0.163	0.274	0.353	1.000											
	Shy, under confident in task	-0.142	0.010	-0.023	-0.069	-0.127	-0.095	-0.063	-0.020	-0.143	-0.012	1.000										
	Set rules/expectations	0.192	0.090	0.007	0.233	0.196	0.158	0.210	0.307	0.242	0.142	-0.025	1.000									
	Broke task in small steps	0.139	0.092	0.032	0.210	0.293	0.248	0.171	0.453	0.280	0.286	-0.045	0.379	1.000								
	Turn task into game	0.223	0.027	0.061	0.134	0.353	0.286	0.218	0.326	0.342	0.331	-0.140	0.290	0.428	1.000							
	Used descriptive words	0.158	0.028	0.030	0.180	0.152	0.103	0.310	0.197	0.283	0.168	-0.063	0.175	0.202	0.260	1.000						
During	Told child name of things/people	0.056	0.082	0.000	0.078	0.064	0.096	0.087	0.128	0.094	0.050	0.040	0.129	0.081	0.065	0.200	1.000					
interview,	Praised	0.307	0.128	0.044	0.107	0.183	0.153	0.136	0.123	0.191	0.128	-0.077	0.145	0.161	0.157	0.121	0.265	1.000				
	Conveyed positive feelings	0.098	0.064	0.019	0.036	0.200	0.207	0.183	0.118	0.128	0.201	-0.022	0.104	0.220	0.187	0.156	0.262	0.390	1.000			
	Caressed/kissed	0.073	0.265	0.050	0.166	0.068	0.061	0.074	0.123	0.067	0.064	0.028	0.031	0.153	0.077	0.081	0.154	0.178	0.218	1.000		
	Kept child in visual range	0.025	0.089	0.011	0.063	0.155	0.187	0.117	0.097	0.129	0.210	0.034	0.067	0.226	0.139	0.107	0.158	0.246	0.395	0.201	1.000	
	Responded verbally to vocalisations	0.100	0.015	0.050	-0.024	0.151	0.192	0.097	0.111	0.179	0.211	-0.010	0.079	0.132	0.237	0.133	0.179	0.226	0.373	0.154	0.393	1.000
	Cronbach alpha	**0.806**																				

Figure B15: Correlation between items used to measure negative parenting behaviours

	Heada-ches	Poor appetite	Sleep badly	Easily frighte-ned	Hands shake	Feel nervous	Poor digestion	Trouble thinking	Unhappy	Cry more than usual	Difficult to enjoy daily	Difficult to decide	Daily work suffering	Unable play useful part	Lost interest in things	Worth-less person	Thoughts ending life	Tired all the time	Uncom-fortable feelings	Easily tired
Headaches	1.000																			
Poor appetite	0.442	1.000																		
Sleep badly	0.305	0.481	1.000																	
Easily frightened	0.242	0.329	0.380	1.000																
Hands shake	0.183	0.233	0.201	0.228	1.000															
Feel nervous	0.174	0.320	0.301	0.442	0.318	1.000														
Poor digestion	0.142	0.246	0.269	0.234	0.252	0.288	1.000													
Trouble thinking	0.120	0.226	0.276	0.339	0.173	0.351	0.335	1.000												
Unhappy	0.153	0.295	0.304	0.309	0.220	0.350	0.271	0.366	1.000											
Cry more than usual	0.124	0.198	0.206	0.290	0.204	0.321	0.243	0.318	0.397	1.000										
Difficult to enjoy daily	0.136	0.220	0.267	0.285	0.176	0.340	0.268	0.323	0.336	0.345	1.000									
Difficult to decide	0.103	0.221	0.248	0.274	0.159	0.314	0.302	0.404	0.351	0.356	0.478	1.000								
Daily work suffering	0.167	0.242	0.246	0.238	0.153	0.257	0.208	0.258	0.273	0.252	0.410	0.324	1.000							
Unable play useful part	0.102	0.201	0.198	0.243	0.164	0.283	0.267	0.274	0.269	0.257	0.366	0.383	0.304	1.000						
Lost interest in things	0.118	0.256	0.219	0.265	0.228	0.296	0.247	0.284	0.359	0.329	0.342	0.383	0.292	0.410	1.000					
Worthless person	0.080	0.190	0.199	0.251	0.194	0.304	0.241	0.261	0.302	0.295	0.309	0.303	0.229	0.339	0.404	1.000				
Thoughts ending life	0.100	0.133	0.139	0.179	0.139	0.187	0.182	0.187	0.236	0.280	0.186	0.232	0.155	0.158	0.219	0.286	1.000			
Tired all the time	0.158	0.239	0.185	0.230	0.184	0.260	0.203	0.232	0.291	0.229	0.297	0.283	0.345	0.256	0.249	0.293	0.237	1.000		
Uncomfortable feelings	0.167	0.248	0.234	0.226	0.201	0.250	0.331	0.285	0.268	0.244	0.294	0.302	0.243	0.298	0.316	0.262	0.178	0.393	1.000	
Easily tired	0.136	0.204	0.199	0.245	0.187	0.254	0.192	0.237	0.278	0.199	0.281	0.236	0.281	0.207	0.241	0.219	0.196	0.608	0.397	1.000
Cronbach alpha	**0.876**																			

Figure B16: Correlation between items used to measure primary carer's depressive symptoms (SRQ-20)

	Satisfied	No good at all	Qualities	Do things	Not much proud	Useless	Worthy	Low self-respect	Feel failure	Positive attitude
Satisfied	1.000									
No good at all	0.056	1.000								
Qualities	0.278	0.038	1.000							
Do things	0.263	0.028	0.344	1.000						
Not much proud	-0.056	0.242	-0.132	-0.085	1.000					
Useless	0.069	0.331	0.048	0.012	0.317	1.000				
Worthy	0.287	-0.013	0.249	0.342	-0.070	0.003	1.000			
Low self-respect	-0.127	0.156	-0.151	-0.129	0.243	0.183	-0.137	1.000		
Feel failure	-0.002	0.170	0.009	0.020	0.230	0.359	-0.009	0.081	1.000	
Positive attitude	0.248	0.062	0.319	0.248	-0.127	0.063	0.211	-0.159	-0.006	1.000
Cronbach alpha	**0.597**									

Figure B17: Correlation between items used to measure primary carer's self-esteem (Rosenberg scale)

Test	p-value	Additional parameters freed
A - Cognitive skills		
Metric vs Configural	0.127	
Scalar vs Configural	0.096	
Partial Scalar vs Configural	0.405	Intercepts IDELA8
Strict vs Configural	0.192	
B - Emergent Literacy		
Metric vs Configural	0.118	
Scalar vs Configural	0.130	
Strict vs Configural	0.023	
C - Emergent Numeracy		
Metric vs Configural	0.277	
Scalar vs Configural	0.045	
Partial Scalar vs Configural	0.316	Intercepts IDELA8
Strict vs Configural	0.256	
D- Executive Function		
Metric vs Configural	0.815	
Scalar vs Configural	0.748	
Strict vs Configural	0.411	

Figure B18: Tests of measurement invariance for cognitive measures

Test	p-value	Additional parameters freed
A - Emotional awareness		
Metric vs Configural	0.881	
Scalar vs Configural	0.416	
Strict vs Configural	0.449	
B. Emotional and behavioural problems		
Metric vs Configural	0.016	
Partial Metric vs Configural	0.636	FL SDQ stealing
Partial Scalar vs Configural	0.018	Intercepts SDQ stealking
Partial Scalar vs Configural	0.181	Intercepts SDQ popular, attends
Strict vs Configural	0.000	
C - Externalising behaviours		
Metric vs Configural	0.000	
Partial Metric vs Configural	0.000	FL SDQ steals, attends, obeys, fights, fidgety, reflect, lies, and restless
Partial Scalar vs Configural	0.602	Intercepts SDQ steals, attends, obeys, fights, fidgety, reflect, lies, and restless
Strict vs Configural	0.000	
D. Internalising behaviours		
Metric vs Configural	0.451	
Scalar vs Configural	0.149	
Strict vs Configural	0.131	

Figure B19: Tests of measurement invariance for socio-emotional measures

Test	p-value	Additional parameters freed
A - Knowledge of Infant Development Inventory		
Metric vs Configural	0.333	
Scalar vs Configural	0.323	
Strict vs Configural	0.000	
B - Positive parenting behaviours		
Metric vs Configural	0.002	
Partial Metric vs Configural	0.141	FL rewards, feedback, kiss, spontaneous praise, focus child when lose Interceptserest
Scalar vs Configural	0.000	Intercepts rewards, feedback, kiss, spontaneous praise, focus child when lose Interceptserest
Partial Scalar vs Configural	0.176	Intercepts signs of affection, convey positive feelings, keep child in visual range
Strict vs Configural	0.000	
C - Negative parenting behaviours		
Metric vs Configural	0.000	
Partial Metric vs Configural	0.148	FL hit, spank, called names, hit, shouted at (activity), shouted at (Interceptserview), anger, Interceptsrusive behaviour, shook (Interceptserview), gave something to do, slap, disinterested, shook (activity)
Partial Scalar vs Configural	0.271	Intercepts hit, spank, called names, hit, shouted at (activity), shouted at (Interceptserview), anger, Interceptsrusive behaviour, shook (Interceptserview), gave something to do, slap, disinterested, shook (activity)
Strict vs Configural	0.000	

Figure B20: Tests of measurement invariance for measures of parental knowledge and parenting behaviours

Test	p-value	Additional parameters freed
A -Depressive Symptoms		
Metric vs Configural	0.298	
Scalar vs Configural	0.000	
Partial Scalar vs Configural	0.224	Intercepts for always tired, poor digestion, difficulty daily activities, frightened, sleep problems, tired easily
Strict vs Configural	0.492	
B - Self-esteem		
Metric vs Configural	0.630	
Scalar vs Configural	0.016	
Partial Scalar vs Configural	0.315	Intercepts for proud
Strict vs Configural	0.369	

Figure B21: Tests of measurement invariance for measures of primary carer's mental health and self-esteem

	Cognitive	Emerging Numeracy	Behavioural Problems	Externalising Problems	Positive Behaviours	Negative Behaviours	Depressive Symptoms	Self-esteem
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment among non-PMs	0.095	0.071	-0.079	-0.114	0.052	-0.112	0.068	-0.041
	(0.056)	(0.053)	(0.055)	(0.044)	(0.074)	(0.104)	(0.074)	(0.092)
p-val	*0.095*	*0.190*	*0.153*	*0.011*	*0.481*	*0.284*	*0.364*	*0.653*
Treatment among PMs	0.124	0.136	-0.337	-0.401	0.327	-0.007	0.048	0.135
	(0.079)	(0.088)	(0.119)	(0.115)	(0.119)	(0.220)	(0.124)	(0.107)
p-val	*0.120*	*0.126*	*0.006*	*0.001*	*0.007*	*0.973*	*0.701*	*0.212*
Difference in treatment	0.029	0.065	-0.258	-0.287	0.275	0.105	-0.020	0.176
	(0.071)	(0.087)	(0.101)	(0.105)	(0.097)	(0.194)	(0.108)	(0.113)
p-val	*0.683*	*0.455*	*0.013*	*0.008*	*0.006*	*0.591*	*0.854*	*0.125*
Control Mean Endline non-PMs	0.013	0.016	-0.060	-0.077	0.006	-0.031	0.005	-0.020
Control Mean Endline PMs	0.160	0.169	-0.040	-0.064	0.258	0.088	-0.009	0.206
Baseline outcome	Yes	No	Yes	Yes	No	No	No	No
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2151	2151	2151	2151	2151	2,151	2151	2,151

Figure B22: Impacts on measures allowing for partial measurement invariance, by PM