USC Dornsife
Center for Economic and Social Research

USC Schaeffer
Leonard D. Schaeffer Center for Health Policy & Economics

Group Meetings and Boosters to Sustain Early Impacts on Child Development: Experimental Evidence from Kenya

Italo Lopez Garcia, Jill E. Luoto, Frances E. Aboud, Lia C.H. Fernald

Paper No: 2023-007

CESR-SCHAEFFER WORKING PAPER SERIES

cesr.usc.edu **healthpolicy.usc.edu**

Group Meetings and Boosters to Sustain Early Impacts on Child Development: Experimental Evidence from Kenya.[1]

Italo Lopez Garcia*
Center of Economic and Social Research
University of Southern California

Jill E. Luoto*
Center of Economic and Social Research
University of Southern California

Frances E. Aboud
Department of Psychology
McGill University

Lia C.H. Fernald
School of Public Health
University of California, Berkeley

Abstract: We present results two years after the end of a group-based parenting intervention tested in a cluster randomized control trial in rural Kenya. The original program consisted of 16 fortnightly village-based sessions over 8 months and had large positive impacts on children's cognition and parenting behaviors immediately after its end. Over the next two years, a random half of intervention villages received a light-touch "booster" intervention to offer continued yet less intensive program support. With and without the booster extension, early program impacts were sustained two years later, albeit smaller in magnitude. Boosters had a small positive added value on parenting behaviors and children's socioemotional development, despite the interruption of COVID-19 to their delivery. Sustained impacts on children's development were strongly mediated by improvements in parenting behaviors, disadvantaged families accrued the largest benefits, and two years later our program remains one of the most cost-effective and potentially scalable programs globally to date. These results point to encouraging paths forward for maximizing the reach and longer-term effectiveness of early childhood development programs to improve child development in low-resource remote settings. (JEL No: H43, I10, I20, I38, O15)

Keywords: parenting intervention, parenting behaviors, early child development, group-based delivery, rural Kenya

---

[1] We thank participants in several seminars and conferences for comments. This work is supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development with award number R01HD090045. A special thanks goes to Edith Alu, Alie Eleveld, Alex Mwaki, Ronald Otieno, Aloyce Odhiambo, Ursula Liona, Fredrick Mseveni, George Simbiri, Joab Ochieng, the community health volunteers, and the entire team at the Safe Water and AIDS Project in Kenya for invaluable support in the carrying out of this study. We are grateful to Douglas Newball Ramirez for invaluable research assistance. Finally, our greatest thanks go to all the parents and children who participated in this study. All errors and omissions are our own.

*Co-first authors

1 Introduction

More than 250 million children under age 5 (43%) from low and middle-income countries (LMICs) are predicted not to reach their developmental potential as adults due to extreme poverty and related causes (Black et al. 2017) Children growing up in poverty are exposed to risk factors that can compromise their early childhood development (ECD), and early deficits can accumulate over the life cycle, with the potential for severe negative impacts on economic and health outcomes as adults (Heckman and Mosso 2014; Almond, Currie, and Duque 2018). Improving the early environments of disadvantaged children can be both an efficient and equitable means to improve life-course outcomes (Heckman 2006; Knudsen et al. 2007). Parenting behaviors play a critical role during sensitive periods of early childhood (Cunha, Heckman, and Schennach 2010; Fiorini and Keane 2014; Attanasio et al. 2020), and the negative effects of poverty on parenting are well-established (Carneiro, Meghir, and Parey 2013; Kalil and Ryan 2020). There is a large research gap relating to how best to design and scale ECD parenting interventions to promote sustained improvements in children's development over time and with the potential to break the intergenerational transmission of poverty, particularly in LMIC settings (Britto et al. 2017).

Parenting interventions featuring parent-child responsive stimulation and improved caregiving can effectively improve ECD outcomes in LMIC settings, at least in the short term (Jeong et al. 2021). However, to date, the vast majority of these interventions have been delivered via individual home visits, which can be a prohibitively expensive and time-consuming model to deliver services at scale in resource-poor settings (Aboud and Yousafzai 2015). Accordingly, very few evidence-based programs have been scaled to the regional or national level (Radner et al. 2018; Tomlinson et al.,2018). Furthermore, the very few home visiting programs that have examined their sustained effects two or more years later also find that early impacts tend to fade out over time (Jeong, Pitchik, and Fink 2021). In contrast, community-based group meetings could be a more cost-effective and scalable model to deliver parenting programs for resource-poor LMIC settings, but have been relatively less examined. Groups can provide social support and allow for peer-to-peer learning, but may be comparatively weak in providing opportunities to actively practice the behaviors and overcome barriers to behavior change (Yousafzai and Aboud 2014). In spite of these potential limitations, recent evidence shows that group meetings are at least as effective as, and more cost-effective than, individual home visits to improve children's developmental outcomes in the short-term (Aboud and Yousafzai 2015; Grantham-McGregor et al. 2020). However, evidence on their ability to sustain impacts over time is still very limited. More evidence is needed on parenting interventions that are affordable, potentially scalable, and that can achieve sustained impacts over time on child development in impoverished and rural settings such as Sub-Saharan Africa.

In this paper, we examine whether parenting interventions that are explicitly designed to be affordable and potentially scalable for low-resource, rural settings can improve parenting behaviors and ECD outcomes, and with impacts that are sustained over time. We also test the role of extending program support after the end of the main intervention to help sustain improvements in parent and child outcomes over time. These questions are even more important for low-resource remote settings such as rural Kenya, where a majority (56%) of children under age 5 are growing up in poverty and local health services tasked with delivering nurturing care services in their communities are severely underfunded (Lu, Black, and Richter 2016; Abboah-Offei et al. 2022).

We evaluate the impacts two years after the end of *Msingi Bora (*Swahili for *"Good Foundation")*, a group-based parenting intervention delivered fortnightly over 8 months that was shown to improve child developmental outcomes and parenting behaviors measured immediately after its completion. Over the following two years, we introduced a light-touch "booster" intervention in half of the original intervention villages to test the added value of offering continued program support to help sustain early impacts longer-term.

The main intervention was tested in a 3-arm cluster randomized trial (cRCT) across 60 villages and 1152 households with children aged 6-24 months at baseline in rural western Kenya. Twenty villages were randomly assigned to a group-only delivery model receiving only group sessions within villages, another 20 to a mixed-delivery model receiving a combination of group sessions and personalized home visits, and the remaining 20 villages comprised a control group. A total of 16 fortnightly sessions over an 8-month period were delivered by trained community health volunteers (CHV) to mother-child dyads in both treatment arms.[2] The goal of this original experiment was to test the most effective – and cost-effective - delivery model for an ECD responsive parenting intervention that could be potentially scalable for rural, resource-poor settings such as western Kenya. The short-term results from a first follow-up survey collected at the end of the main intervention, when children were an average of 26 months old, showed the two delivery models were highly effective and cost-effective at improving measures of parenting behaviors and child development (Luoto et al. 2020; Lopez Garcia, Saya, and Luoto 2021). Building on these findings, we now present results from a follow-up survey conducted two years after the end of the original 8-month intervention, when children were 3.5-5 years old, to test its sustained impacts over time. In addition, we test the added value of hosting booster group meetings every other month in half of intervention villages during the two years between follow-up surveys. While the onset of the global COVID-19 pandemic interrupted the boosters' implementation, a total of nine booster sessions were held in selected villages during this period. This design results in four distinct intervention groups by the time of this two-year survey, each comprised of 10 villages: groups without boosters, groups with boosters, mixed without boosters, and mixed with boosters.

The short-term impacts of *Msingi Bora* on individual measures of children's development and parenting practices were previously published (Luoto et al 2020). Here, we estimate latent factors of children's skills and parental outcomes that correct for measurement error in individual measures and facilitate comparisons with impacts two years later estimated on a different set of individual measures (as children aged out of the original scales). Using our new latent factors, we are able to replicate the large short-term impacts on children's cognitive development from both delivery models, with the group-only arm slightly outperforming the mixed-delivery arm (0.5 SD vs 0.3 SD, respectively, p=0.085 on one-sided test). We also find positive short-term

[2] Within this 8-month intervention period, half of villages randomized to an intervention arm also invited fathers to the sessions and held 4 separate father-only sessions. Very few fathers attended sessions and inviting them had no additive short-term impacts on children or any other outcome (Luoto et al. 2020). The father randomization was dropped from consideration after the 8-month program and is not further discussed in this manuscript.

impacts on maternal stimulation and knowledge of child development that were statistically the same across delivery models, and no impacts on maternal wellbeing.

Two years later, our results show that the group-only delivery model, without the booster extension, had sustained impacts on children's cognitive and socioemotional development of about 0.2 SD, with similarly-sized impacts on maternal stimulation and knowledge. These medium-term impacts represent sizable declines of 59%-77% on children's cognition and maternal stimulation from two years prior, and much of these declines are due to families in the control group "catching up," particularly among better-off families. The addition of boosters to group-only visits resulted in slightly larger mean impacts across all outcomes, but we cannot reject equality of outcomes with villages that did not receive boosters. For villages that received the mixed-delivery model but no supplemental boosters, all short-term impacts on child and parental outcomes faded from significance after two years. The combination of the mixed-delivery model with subsequent boosters did lead to positive impacts on children's socioemotional scores, as well as maternal stimulation and knowledge two years later, but not on children's cognition. Finally, pooling group-only and mixed-delivery villages together, we find that boosters had a small but statistically significant additive impact on children's socioemotional development and maternal stimulation relative to villages that did not receive the boosters.

To provide a more comprehensive examination of the program benefits to inform policy, we perform three additional analyses after two years: an examination of heterogeneous treatment effects to understand who benefited most from the program, a mediation analysis to explore the mediating pathways explaining sustained program impacts, and a benefit-cost analysis to assess the potential long-term net benefits of scaling the program. Since impacts in the mixed-delivery arm largely faded after two years, these additional analyses only focus on the group-only arm.

An important goal of ECD parenting interventions is to reduce early gaps in children's development that stem from socioeconomic disparities. We use the Sorted Partial Effects method (Chernozhukov et al., 2018) to flexibly examine heterogeneous effects and find that children with the largest gains in cognition after two years came from households that were 1.6 SD poorer on average (using our standardized baseline wealth index) than children who had the smallest cognitive impacts after two years. We see a similar pattern for maternal stimulation behaviors, both with and without boosters, but not for children's socio-emotional scores.

In terms of exploring the underlying pathways explaining our results after two years, both current impacts on maternal stimulation behaviors and knowledge, as well as past program impacts on children's cognitive development, strongly mediate the two-year impacts on children's cognition for group-only villages, with or without the booster extension. For example, an exploratory mediation analysis suggests that the combined effect of these three mediators explains up to 78% of the observed impacts after two years on children's cognition in villages that did not receive subsequent boosters. In the same villages, these same three mediators explain up to 61% of the sustained impacts on children's socioemotional outcomes. Notably, the fact that past program impacts on children's cognition have independent mediating roles on current impacts on children's cognitive and socioemotional outcomes two years later suggests the presence of strong dynamic complementarities in children's skills over time.

Finally, we find that our group-only *Msingi Bora* intervention, with or without subsequent boosters, remains highly cost-effective two years later. The overall cost per child without boosters is $119 in 2020 US$, and $157 including the booster extension. Adjusting for inflation and translating to purchasing power parity (PPP) terms, and combined with our sustained impacts after two years, *Msingi Bora* is one of the most, if not the most, cost-effective parenting interventions to date from a rural, resource-poor LMIC setting. Even after taking into account societal costs such as the private opportunity costs for both mothers and CHVs to attend or host the sessions, the smaller program impacts after two years, and the disruption of the booster sessions by the COVID-19 pandemic, our projected benefits outweigh costs by a factor of 6.9-7.5, making the program highly sustainable from a social planner's perspective.

Our study makes four main contributions to the literature. First, we demonstrate the sustained effectiveness and cost-effectiveness of a purely group-based ECD parenting intervention in a LMIC setting two years after the end of the main intervention period. Though other group-based parenting programs have recently demonstrated sustained impacts two or more years after the end of their programs, these are either from high-income countries (Carneiro et al. 2023), combined group visits with more expensive home visits delivered intensively over two years (Yousafzai et al. 2016; Meghir et al. 2023)[3], or utilized a low-quality measure of child development based on maternal report that is potentially biased (Justino et al. 2022).[4]

To the best of our knowledge, our study is also the first formal test of a program extension to combat the common problem of fade-out affecting parenting behavior change programs. Though our "light-touch" booster extension had only small and marginal impacts on children's socioemotional development and maternal stimulation behaviors, these results are encouraging in light of the global COVID-19 pandemic that unfolded at the same time of the booster extension. The pandemic severely disrupted their implementation, and limited our ability to provide in-person trainings to CHV delivery agents as well as monitor their delivery quality, factors that were key to the success of the main trial (Luoto et al. 2021). We also see this as the key factor behind the inability of booster sessions to further improve children's cognition beyond that achieved from the original intervention, in addition to the fact that the more-intensive program had already increased children's cognition by sizable amounts.

A third contribution of our paper relates to the ability of our group-based parenting program to close early developmental gaps across socioeconomic groups by benefiting more children from the poorest households, which can be a highly cost-effective policy tool to tackle long-term inequality among the most deprived populations in rural Sub-Saharan Africa.

---

[3] Meghir et al. (2023) find sustained impacts for the Reach-up parenting program in India, but only after 15 months, and combined across their group and home visit delivery models, making the sustainability of their group-only delivery model effectively unknown. This program is also significantly more expensive than *Msingi Bora* (Lopez Garcia, Saya, and Luoto 2021)

[4] The Justino et al. (2022) study from rural Rwanda utilized the Ages and Stages Questionnaire 3rd edition (ASQ-3), which previous studies have found to have poor concurrent validity with measures based on direct assessment of children (Rubio-Codina et al. 2016; Duggan et al. 2023; Veldhuizen et al. 2015; Yue et al. 2019) even in the context of another home visiting intervention also from Rwanda.(Jensen et al. 2021). Other potential concerns about the study's design include it involved just 9 villages, suffered from contamination into the control villages, and the only intervention model demonstrating sustained impacts featured a home visit, whereas their purely group-based model faded from significance after two years.

Finally, findings from our mediation analysis highlight the huge importance of achieving sustained changes in parenting behaviors to help sustain the effects of parenting interventions on children's development for a longer period of time, as well as the importance of dynamic complementarities of parental investments in children over time (Cunha, Heckman, and Schennach 2010). Often hypothesized, such dynamics are not often demonstrated empirically due to a lack of suitable data. Though not fully causal, our study is among the very few to perform a dynamic mediation analysis relating parental inputs and child outcomes using three or more waves of data in LMICs (Attanasio, Meghir, and Nix 2020; Attanasio et al. 2017), and the first to find strong suggestive evidence of dynamic complementarities in the context of an experiment from Sub-Saharan Africa.

The rest of this paper proceeds as follows. Section 2 describes the *Msingi Bora* program, including the Booster extension. Section 3 presents the experimental design, child and parental measures, and basic tests of balance and study attrition. Section 4 presents our empirical strategy and Section 5 presents our main results. In Section 6 we present complementary analyses including heterogeneous treatment effects, a mediation analysis and an examination of program benefits and costs. In Section 7 we discuss our results and conclude.

2. Background: The *Msingi Bora* Intervention

The *Msingi Bora* intervention was adapted from previous successful parenting trials from our team in Uganda and Bangladesh (Singla, Kumbakumba, and Aboud 2015; Aboud and Akhter 2011) and expanded to include more activities around responsive play and talk with children (Luoto et al. 2020). *Msingi Bora* is based on Social Learning Theory (Bandura 1995; 1986) and aims to teach parents how to implement responsive stimulation practices as recommended in the Nurturing Care Framework (World Health Organization, 2018). It emphasizes active learning by parents through demonstration, coaching and practice, and uses a structured curriculum to teach parents responsive activities using materials freely available at home as well as two-way responsive talk using available pictures from home. Similar to other child development curricula, *Msingi Bora* directs new learning at the parent, where parents are taught to be responsive by adjusting the activities to their child's abilities as they get coached and as children grow.[5] The program also promotes peer-to-peer learning through group discussions of common barriers and ways to resolve them, as well as its use of a "buddy system," in which smaller groups of 2-3

[5] Two of the most widely-known ECD parenting curricula in use in LMIC settings are the Reach Up curriculum (Heckman 2023; S. M. Grantham-McGregor et al. 1991; Jervis et al. 2023). and the UNICEF/WHO Care for Child Development (CCD) program (Ahun et al. 2023). Similar to *Msingi Bora*, Reach Up focuses on teaching age-specific skills to the caregiver and child - in one-on-one visits, or more recently in group-based programs - and also uses a structured manual that outlines a specific skill at a specific age point (Jervis et al. 2023). Unlike *Msingi Bora*, Reach Up does not cover topics on the key messages of nutrition, hygiene, or love and respect in the family, and usually loans the families the play materials to be used at home for practice. CCD is based on attachment theory (Bretherton 1992), and focuses on counseling the parent on age-specific stimulation, through a series of counselling cards that illustrate how caregivers can play and talk with their child. The CCD curriculum is significantly less structured, with just seven counselling cards on play, communication, and nutrition activities -- five for home visits from birth to 11 months, one for 1 year of age and one for 2 years. Parents are asked how they currently play with and talk to their child. Depending on their answer, new activities may be suggested. However, new play and talk is not consistently demonstrated and coached, as with the other two programs.

parents are encouraged to support each other between sessions by meeting to practice the activities from the previous session and complete assigned homework. The curriculum focused on five key practices: responsive play, responsive talk, hygiene, nutrition, and love and respect within the family. Mothers received an illustrated brochure of the five key practices and one local picture book (in a later session), but otherwise collected their own home-available playthings to include in a playbag. Over time, parents were encouraged to add new play materials in the play bag to support more complex stimulation activities as children aged.

2.1 Original Intensive *Msingi Bora*

The original *Msingi Bora* program consisted of 16 sessions delivered every other week in intervention villages over an 8-month period.[6] Six sessions were piloted in April-June 2018 in six villages not included in the main trial. The finalized curriculum included session-specific activities and materials, with Luo or Swahili and English manuals for each CHV delivery agent. CHVs were existing personnel from Kenya's rural health system and under local policy are already tasked with the delivery of nurturing care services in their communities through home visits (Kenya Ministry of Health 2020). The local Kenyan non-governmental organization Safe Water and AIDS Project (SWAP) provided supervisory and training capacity in *Msingi Bora*'s implementation under a train the trainers model.

Group-based sessions took place in local community centers or churches within villages. CHVs in villages assigned to an intervention arm were paid a monthly stipend for their duties, according to local custom, and underwent eight days of intensive training in November 2018 covering sessions 1 to 8, and another eight days in April 2019 covering sessions 9 to 16. Monthly one-day refresher trainings were performed in each sub-county for that month's sessions. Overall, CHVs received a total of 21 days of training that included intensive practice with families from villages that were not part of the study. Group sessions lasted an average of 90 minutes and 89% were monitored by trained SWAP supervisors who rated CHVs on items such as facilitating discussion, coaching parents, answering questions, as well as overall session quality and engagement.

Mothers and their children aged 6-24 months at the time of a baseline survey were invited to attend all 16 sessions, and received a small gift for attendance (e.g., a small bar of soap ($0.15)). Every fourth session served as a review session, for which households in the group-only arm continued with group meetings while households in the mixed-delivery arm received individual home visits from their same CHV and during the same weeks that a group review session was held in group-only villages. During these home visits, CHVs delivered review messages identical to those in the group reviews, but the focus was personalized on that family. Households in villages assigned to the control group did not receive any intervention besides information about child feeding during the baseline survey.

2.2 Booster Extension of *Msingi Bora*

[6] The curriculum for the main intervention is publicly available at https://www.mcgill.ca/psychology/files/psychology/msingi_bora_responsive_parenting_manual.pdf

The intensive phase of *Msingi Bora*'s 16 sessions was completed in July 2019, and a first follow-up survey on its short-term impacts was conducted across all 60 villages in August-October 2019. Beginning in November 2019, we extended the curriculum in the form of group "booster" sessions intended to be held every other month within selected villages. The goal of the booster extension was to test the value-added of offering "light touch" continued program support over two years to help sustain improvements in parenting behaviors and child outcomes. The boosters' curriculum was designed to build on the original 16 sessions and added new and more advanced strategies for responsive play and talk as children grew more capable, as well as introduced positive disciplinary practices as children grew older. Each booster session lasted an average of 90 minutes and reinforced the five key messages of the original program while introducing the new curriculum and topics in the form of group discussions, skits, and guided practice between mothers and children for any new activities.[7]

These boosters were designed to be held in groups, based on short-term impacts from the 2019 survey revealing the group-only model outperforming the mixed-delivery model across a range of child and parental outcomes, as well as findings from qualitative exit interviews with mothers after the first 16 sessions in which mothers expressed a preference for group sessions over home visits (Luoto et al. 2021). The onset of the global COVID-19 pandemic in March 2020 caused a 6-month delay between booster sessions 2 and 3 and forced us to adjust the design of remaining booster sessions to meet local public health guidance.[8] A total of nine booster sessions were able to be held between November 2019 and August 2021.

2.3 Program take-up

Attendance to all *Msingi Bora* sessions was voluntary. During the first intensive phase with fortnightly sessions, almost all invited families (97%) participated in at least one session, and attendance averaged 69% across both delivery models. The median mother-child dyad attended 13 of the 16 sessions. For the booster extension, attendance was similar (70%), and the median mother-child dyad attended 8 of 9 boosters held every other month.

3 The Study Design and Data

The *Msingi Bora* trial was implemented in the sub-counties of East Rachuonyo, South Rachuonyo, and Sabatia in Western Kenya. These predominantly rural areas are characterized by high rates of poverty and child mortality (DHS 2022). Sabatia's population is predominantly from the Luhya tribe and speaks Luhya and Swahili. East and South Rachuonyo are predominantly Luo and speak Luo. These subcounties were chosen because they were within driving distance of SWAP's headquarters in Kisumu town. The majority of villagers are subsistence farmers or unskilled informal workers.

---

[7] The curriculum for the booster extension is publicly available at https://www.mcgill.ca/psychology/files/psychology/msingi_bora_responsive_parenting_booster_sessions.pdf

[8] The primary changes due to COVID-19 affecting the boosters' delivery included that all trainings of SWAP's main training team were done remotely over Zoom versus in-person, booster sessions were moved outside, participants were provided with masks, and groups were subdivided to reduce group sizes from an average of 12-13 mothers to 4-5. This meant that CHVs delivered each booster session more than once per village on average.

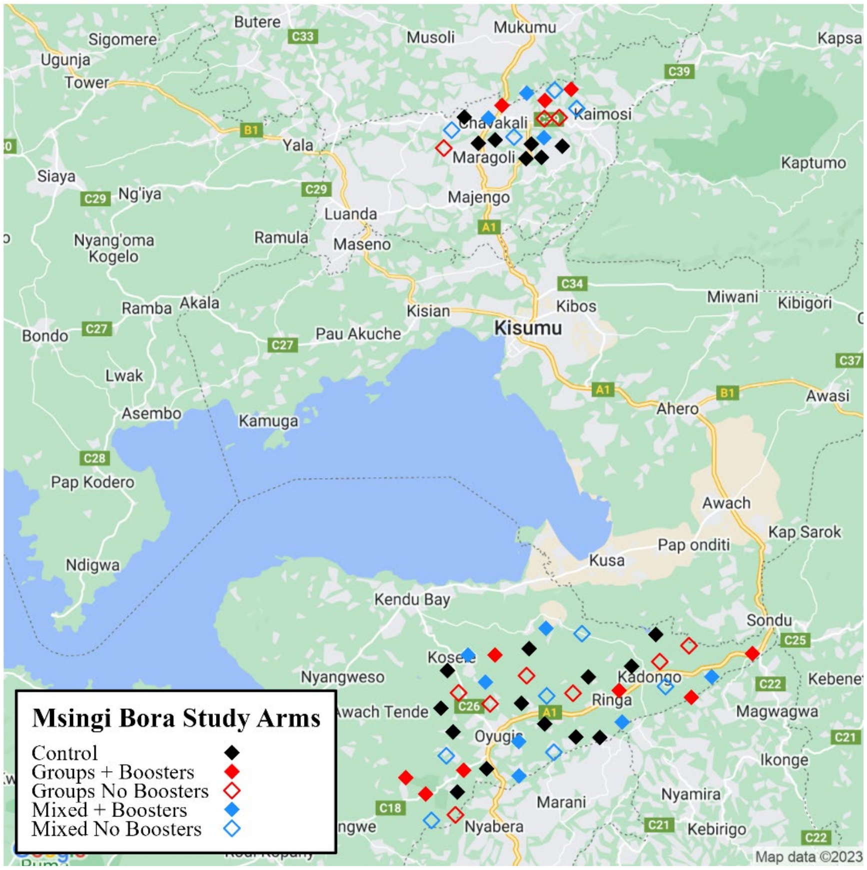
3.1 Recruitment and Randomization

To recruit villages, SWAP first listed all villages across the three study subcounties that were estimated to have at least 20 households with age-eligible children. Stratifying by subcounty and using a computer-generated random number in Stata, 60 villages were randomly sampled for inclusion as long as they maintained a minimum distance (1.5-2.5 kms) from all other sampled villages. Villages made up clusters and were the unit of randomization to minimize risk of contamination within villages. We invited CHVs from the sampled villages to participate in the study and made them aware of the likely time commitment if their village was assigned to an intervention arm. No CHV refused participation.

A team of trained enumerators conducted a census within sampled villages to identify all eligible households and record GPS coordinates to facilitate collection of surveys. Eligible participants were mothers or other female primary caregivers aged 15 and over with a child between 6-24 months at the time of the baseline survey, without signs of severe mental or physical impairment. From this list, we drew a sample of 20 households using a random number generator, with 10 more names on a ranked waitlist in case of study refusal. Very few households refused participation. Immediately after the collection of the baseline survey in Fall 2018, the 60 study villages were randomly assigned to one of three study arms, stratified by subcounty, each 20 villages in size: the group-only delivery model (Arm 1), the mixed-delivery model (Arm 2), and the control group (Arm 3). Arms 1 and 2 received the main intervention with fortnightly sessions over 8 months.

At the end of the original 8-month intervention period, villages assigned to the two intervention arms underwent a secondary randomization procedure to determine those that would receive the booster extension every other month for the next two years, stratified by subcounty and original intervention arm. This secondary randomization effectively created a 2x2 factorial design among the 40 original intervention villages, with 10 villages that each received a combination of group-only or mixed-delivery for the 8-month program, and then with and without group boosters over the subsequent two years. Figure 1 shows the geographical location of the sampled villages and their final randomization status.

Figure 1: Randomization map



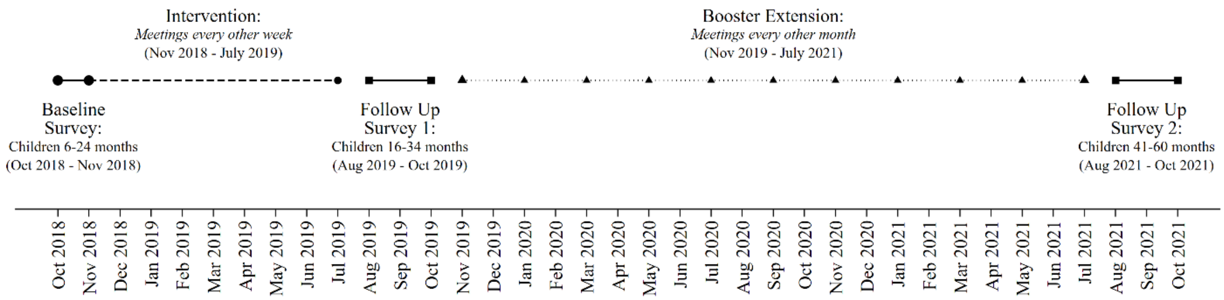
Notes: The figure shows the map of the 2x2 cluster randomization across Sabatia, East Rachuonyo and South Rachuonyo Sub-Counties. The first randomization assigns 60 villages into the group-only arm, the mixed-delivery arm, and the control group. The second randomization pools the 40 treatment villages and assigns them either to receive the supplemental booster sessions or not.

3.2 Surveys and Timeline

Figure 2 shows a timeline of activities for the three-year study. A baseline survey on mothers and children was conducted from October-November 2018. The original intensive *Msingi Bora* intervention took place between late November 2018 and July 2019. A first follow-up survey was fielded from August-October 2019, when children were 16-34 months old. The booster extension was implemented in selected villages (with the exception of the COVID-induced delays noted above) between November 2019 and July 2021, and the two-year follow-up survey was fielded from August-October 2021, when children were 41-60 months (3.5-5 years) old. Survey enumerators were blinded to the household's intervention status at all waves.

Results from the baseline and first follow-up surveys have been analyzed and published in previous manuscripts (Luoto et al. 2020; Lopez Garcia, Saya, and Luoto 2021). The final two-year follow-up survey – and added value of the booster extension - are the focus of this paper, though we compare our results after two years with our short-term results to learn more about the trajectories of program impacts over time.

Figure 2: Study Timeline



Notes: Booster sessions suffered a 6-month delay during onset of the COVID-19 pandemic beginning March 2020.

Child Development Measures

The primary child development outcomes at baseline and the first follow-up survey were assessed using the Bayley Scales of Infant Development third edition (Bayley's III), a commonly used direct child assessment previously adapted and validated in many African countries for children aged up to 42 months (Rademeyer and Jacklin 2013; Hanlon et al. 2016; Pendergast et al. 2018). We used previous adaptations to the Bayley to make it appropriate for our context such as modifying the picture and stimulus booklets with children, attire and objects that would be familiar to rural Kenyan children (Singla, Kumbakumba, and Aboud 2015). At baseline, due to the young ages of children we collected data on just two subscales of the Bayley III, receptive language and cognition; at the first follow-up survey, we also administered the expressive language subtest.

At the two-year follow-up survey, most children had aged out of the Bayley III. In its place, we used different measures based on direct assessment of children including the block-design subtest of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) 3rd Edition to measure cognitive non-verbal reasoning (Raiford and Coalson 2014). For expressive and receptive language, we used Dholuo and Kiswahili versions of The British Picture Vocabulary Scale - III (BPVS III) (Dunn & Dunn, 2009), previously translated and adapted to the local Kenyan context and tested with children from 2-6 years old (Knauer et al. 2019). At this survey wave we also collected measures of children's executive functioning skills by administering four different tasks of the Tangerine EF Touch App, previously tested on Kenyan children aged 3-6 years old (Willoughby et al. 2019). These tasks included one measure of children's working memory (Pick the Picture), one of cognitive flexibility (Something's the Same), and two Stroop measures of inhibitory control (Arrows and Animal Go/No Go). We combine these distinct tasks into a composite index of children's executive functioning for analysis.

Finally, we collected measures of children's behaviors at both follow-up survey waves using a combination of enumerator observation and maternal report. First, we adapted the Wolke Scale (Wolke, Skuse, and Mathisen 1990), in which enumerators observe and rate on a 1-5 scale several dimensions of children's behavior during testing. At the first follow-up survey, enumerators rated seven observed behaviors included approach, emotional tone, activity level, cooperation, vocalization, emotional security, and exploration. At the two-year follow-up survey, enumerators recorded observations on six observed behaviors: emotional tone, activity level, cooperation, emotional security, reflection vs impulsivity, and concentration. Also at the two-year follow-up survey, enumerators administered to mothers the Strengths and Difficulties Questionnaire (SDQ) (Goodman 1997), a 25-item scale measuring children's adaptive and maladaptive behaviors that can be categorized in five sub-scales: hyperactivity, conduct problems, emotional symptoms, peer relationship problems and prosocial behaviors.

Parental measures

Parental stimulation practices were assessed at baseline with the Family Care Indicators (FCI) (Hamadani et al. 2010), a self-reported measure of 12 materials and activities reflecting parent/child interactions. At both rounds of follow-up surveys, we measured stimulation practices and other types of parenting behaviors with the Home Observation for Measurement of the Environment (HOME) inventory, widely considered the gold-standard measure of the quantity and quality of parenting behaviors and stimulation available in the home (Caldwell and Bradley 2003). At the first follow-up survey, we used the 45-item version targeting children under age 3 years that combines 30 items that require the interviewer to observe the behavior directly or certify the availability of play materials reported by the mother, and 15 items based on maternal report. At the two-year follow-up survey, we used the 55-item version intended for children 3-5 years old that similarly combines 30 observed items with 25 items based on maternal-report. Both versions of the HOME include items related to parental engagement with the child in learning/play activities, availability of play materials, and parental warmth and disciplinary practices.

To measure nutrition behaviors, at all three survey waves we collected a child dietary diversity scale where parents were asked to report the foods eaten by the child in the past 24 hours, and quantified to indicate the number of food categories out of 7, following WHO recommendations for an adequate child diet (World Health Organization and UNICEF 2003).

Maternal knowledge of child development at baseline and the first follow-up survey were assessed with a battery of 9 questions about the expected age at which children generally acquire social and cognitive skills such as recognize their mother, understand spoken words, play with objects, counting, reading, or throwing a ball (Singla, Kumbakumba, and Aboud 2015). At the two-year follow-up survey, mothers were similarly asked 8 questions about what age they should begin to teach children things like colors, games to play, reading a book, counting, or sharing with friends.

We also collected a series of maternal measures about their mental health and wellbeing. In all three survey waves, maternal depressive symptoms were assessed with the Center for Epidemiologic Studies Depression Scale (CESD) (Radloff 1977), and perceived social support from family, friends and the community was measured with the Lubben Social Network Scale (LSNS) (Lubben et al. 2006). Beliefs about the importance of their behaviors for their children's development at the first and two-year follow-up surveys were measured with the Parental Cognitions and Conduct Toward the Infant Scale (PACOTIS) (Boivin et al. 2005). At the first follow-up survey, perceived self-efficacy was also measured with the Self-Efficacy for Parenting Tasks Index-Toddler Scale (SEPTI-TS) (Van Rijen et al. 2014), and maternal stress with The Daily Stress Index (DSI) (Almeida, Wethington, and Kessler 2002).

Finally, at each survey wave we collected rich sociodemographic information including family composition, employment, assets, education and housing conditions. At the two-year follow-up survey, we also asked mothers if the focal child was attending preschool.

3.3 Baseline Balance and Descriptives

Table 1 shows basic characteristics of mothers and children at baseline in 2018. Mothers were on average 28.5 years old, with roughly 9 years of education. Just over 60 percent of households had fathers living in the home at baseline, reflecting the frequency of migration especially among men in this part of Kenya. At baseline, the average focal child was 14 months old and was the third-born child within the family. The average household was comprised of roughly 5.6 members. The baseline sample is well-balanced on these characteristics across both the original randomization (Luoto et al., 2020), as well as the secondary booster randomization that came after the first follow-up survey. No statistically significant group differences are found in any baseline characteristics.

Table 1: Baseline characteristics and sample balance

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Control group	Treated without Boosters	Treated with Boosters	Pairwise group differences (p-value)			Joint F-test
				2 vs. 1	3 vs. 1	3 vs. 2	(p-value)
Panel A. Continuous variables							
Mother's age	29.19	28.61	27.86	0.476	0.117	0.348	0.278
	(9.00)	(8.82)	(8.49)				
Mother's education (years)	8.97	8.98	8.75	0.963	0.259	0.305	0.431
	(2.51)	(2.71)	(2.69)				
Household size	5.58	5.69	5.70	0.492	0.542	0.953	0.735
	(1.89)	(2.13)	(2.74)				
Wealth index	-0.06	0.07	-0.01	0.068	0.554	0.408	0.176
	(0.92)	(1.04)	(1.01)				
Child age (months)	14.20	14.42	13.90	0.637	0.493	0.210	0.442
	(4.66)	(4.91)	(4.78)				
Child's Cognitive Factor	0.00	-0.09	-0.06	0.447	0.630	0.827	0.737
	(0.99)	(1.04)	(1.04)				
HOME Stimulation Factor	0.00	-0.10	-0.17	0.352	0.180	0.583	0.344
	(0.98)	(1.03)	(0.97)				
Panel B. Categorical variables							
Father in household (%)	61.01	61.90	62.78	0.824	0.654	0.823	0.904
Female child (%)	51.19	49.21	48.99	0.585	0.597	0.952	0.819
Birth order				0.120	0.450	0.560	
First (%)	21.75	28.84	25.76				0.100
Second or third (%)	21.75	21.16	20.71				0.945

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fourth or more (%)	56.50	50.00	53.54				0.260
Joint F-test (p-value)				0.075	0.171	0.940	
Observations	377	378	396	755	772	773	
Number of villages (clusters)	20	20	20	40	40	40	

Notes: Standard deviations in parentheses. Column 2 contains mean values for households that received some version of the first intensive phase of *Msingi Bora* (group-only or mixed-delivery) but no booster extension. Column 3 contains mean values for households that received both the intensive phase and subsequent boosters. Balance between control group, mixed-delivery and group-only is demonstrated in Table A1 in the Appendix, based on Luoto et al. (2020). For Panel A, columns 4-6 present p-values for t-tests of equality across groups. For Panel B, these columns present p-values for chi2 tests of independence. These p-values were calculated using cluster robust standard errors at the village level. Column 7 presents 3-way joint F-tests of equality across all groups. For Columns 4-6 joint F-tests p-values of equality across variables are also presented.

3.4 Study Attrition

Since our analysis spans three years and the onset of the COVID-19 pandemic, we are concerned about differential attrition resulting from our intervention. In a previous publication we demonstrated that at the first follow-up survey in 2019, we successfully located 1070 (93%) of the enrolled children and mothers from baseline with no differences across study arms when comparing group-only, mixed-delivery, and control arms (6·6%, 8·0%, and 6·8%; p-value of joint F-test =0·76) (Luoto et al 2020). At the two-year follow-up survey in Fall 2021, we were able to track a total of 944 mother-child dyads to complete all child assessments and the maternal survey. Of those, 25 had previously attrited at the first follow-up survey in 2019, only to re-enter our sample at the time of the two-year follow-up survey in 2021. We include these 25 dyads in our final sample, though results are equivalent if we exclude them from our analysis. This results in an overall three-year attrition rate of 18 percent, which is consistent with rates found in similar ECD programs in Sub-Saharan Africa (e.g., Justino et al. 2022). Most of those families who had attrited at either survey wave had moved away, at least temporarily. In this area of Kenya, frequent, temporary migration is common, often as young people search for work opportunities or due to marriages forming or dissolving (Beguy, Bocquier, and Zulu 2010).

Table 2 shows the correlates of attrition at the two-year follow-up survey. Column 1 shows that being assigned to a treatment arm as part of the original intervention, and being assigned to subsequent booster sessions, do not predict attrition. Column 2 shows this result is robust to conditioning on baseline characteristics, though we see that attrition is more likely among younger mothers with a lower parity focal child for the intervention. Column 3 interacts treatment assignment dummies with a suite of predefined baseline characteristics as a test for differential attrition. Though some individual coefficients are significant, the p-value for the joint F-statistic on all interactions at the bottom of this table is jointly insignificant (p=0.343).[9] This result reassures us that any selection bias from attrition is unlikely to be related to intervention status and should not affect the validity of results presented in the paper.

Table 2: Sample Attrition and Intervention Status

	(1)	(2)	(3)

[9] Table A2 in the Appendix shows the full version of Table 2 that includes all of the interaction coefficients in Column 3.

	Basic Model	Adjusted Model	Full Interactions
Treated, no boosters	-0.002	0.008	0.238
	(0.031)	(0.033)	(0.191)
Treated + Boosters	-0.012	-0.010	-0.289
	(0.025)	(0.027)	(0.149)
South Rachuonyo	0.077*	0.057	0.057
	(0.031)	(0.033)	(0.033)
East Rachuonyo	-0.014	-0.032	-0.034
	(0.026)	(0.030)	(0.031)
Mother's age		-0.006***	-0.008***
		(0.001)	(0.002)
Mother's years of schooling		0.004	-0.012
		(0.005)	(0.010)
Wealth index at baseline		-0.022	-0.013
		(0.015)	(0.029)
Child age at baseline (months)		-0.003	-0.006
		(0.002)	(0.004)
Female child (%)		-0.004	0.020
		(0.022)	(0.040)
Birth order 2		-0.034	0.004
		(0.038)	(0.057)
Birth order 3+		-0.100**	-0.093
		(0.031)	(0.051)
Baseline child cognitive ability (age-std.)		-0.010	-0.030
		(0.011)	(0.019)
Interactions Treatments X Covariates	No	No	Yes
Constant	0.161***	0.404***	0.419***
	(0.025)	(0.079)	(0.115)
Observations	1152	1152	1152
F-test P-value joint significance treatment variables	0.884	0.852	0.343

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses clustered at the village. Results from linear probability models predicting attrition at two-year follow-up survey. Treated w/o boosters takes on a value of 1 if a mother-child dyad received the group-only or mixed-delivery intervention during first phase of *Msingi Bora* and no subsequent boosters. Treated + boosters takes on a value of 1 if a dyad received group-only or mixed-delivery interventions followed by the booster extension. Balanced attrition at the first follow-up survey across the two delivery models was demonstrated in Luoto et al. (2020) and is not repeated here.

4 Methods

4.1 The Measurement System

Because our multiple observed measures of parenting behaviors and children's development are highly correlated and suffer from measurement error, we use standard factor analysis models to estimate underlying latent factors of parenting inputs and child skills. This allows us to correct for measurement error of the individual measures and reduces the number of hypotheses to be tested jointly.

Starting with children's outcomes, we estimate latent factors of children's skills using the available measures for each wave, internally standardized using 2-month age bands to reflect inherent differences in child measures due to age. An initial exploratory factor analysis (EFA) of age-standardized test scores suggested the presence of a single cognitive factor at baseline (BL) and the first follow-up survey (FU1). By the time of the second follow-up survey (FU2) when children are 3.5 to 5 years old, an EFA suggests the relevance of two factors, which we label as cognitive and socioemotional. We use these results to estimate a confirmatory factor analysis (CFA) using Generalized Structural Equation Modelling (GSEM) methods on the available measures.[10]

Formally, let $\theta_{i,t}^{k}$ be the latent factor for individual i at survey wave $t = \{0 = BL; 1 = FU1, 2 = FU2\}$ for child skill $k = \{C, S\}$ (cognitive or socioemotional), and let m_{itj} be the observed jth measure for child i at time t. Assuming that observed measures are additively separable functions of the latent factors, we specify the following measurement system for children's outcomes:

$$m_{itj} = \begin{cases} \mu_{tj} + \alpha_{tj}^{C}\theta_{it}^{C} + \eta_{itj} & for \ \ t = 0, 1 \\ \mu_{tj} + \alpha_{tj}^{C}\theta_{it}^{C} + \alpha_{tj}^{S}\theta_{it}^{S} + \eta_{itj} & for \ \ t = 2 \end{cases} \qquad (1)$$

where μ_{tj} is a constant, α_{tj}^{k} is the factor loading for child skill k at time t, and η_{itj} is the measurement error, which is assumed to be mean zero and uncorrelated across measures j. In equation (1), we allow for a non-dedicated measurement system at FU2, where one child measure (the Wolke scale) is allowed to load onto both cognitive and socioemotional factors. This is because the EFA at FU2 suggested that while the cognitive, expressive and receptive language and executive function measures almost exclusively loaded into the cognitive factor and the SDQ subscales loaded onto the socioemotional factor, the Wolke loaded on both factors nearly equally. In contrast, all available scales at earlier waves loaded onto a single cognitive factor so we estimate a dedicated measurement system in the two earlier waves.[11] The fact that we could not separately identify a socioemotional factor at FU1 is consistent with the developmental psychology literature showing that children's abilities become more differentiated with age (Tucker-Drob 2009). Allowing the Wolke measure to proxy for both factors at the second follow-up survey both increases the continuity in measures with the first follow-up survey, and ensures the socioemotional factor is not based solely on SDQ items collected by maternal-report.

In terms of parenting behaviors, we perform a similar estimation method to identify latent factors of parenting behaviors using data from the HOME inventory at the two follow-up waves and the Family Care Indicators (FCI) at baseline. An EFA suggested the presence of a single factor at baseline related to simulation practices, and at least two latent factors at the two follow-up waves. However, in the follow-up waves our CFA fits a model with the two most important

---

[10] While we attempted to estimate the latent factors for the three waves jointly, the CFA only allowed the joint estimation of latent factors at FU1 and FU2; the factor model at BL was estimated separately.

[11] The available child measures include the Bayley cognitive and receptive language subscales at baseline, and the Bayley cognitive, receptive and expressive language subscales as well as the Wolke scale at the first follow-up survey.

factors: a stimulation factor, which comprises items related to play materials, play activities, and the home space and safety, and a warmth factor, which comprises items related to maternal communication, disciplinary behaviors, and affective interactions with the child.

Let $\theta_{i,t}^{l}$ be the latent factor of parenting input $l = \{I, W\}$ (stimulation or warmth). Similar to the specification in equation (1), we estimate a measurement system characterizing these two latent factors of parenting behaviors, but given the large number of items in the HOME, we estimate these latent factors separately for each wave and allow each item to load onto a single index provided the factor loading from a previous EFA was large (above 0.4). This implies the measurement system now is fully dedicated within each wave. Specifically, all 12 items of the FCI at baseline are used to identify our stimulation factor, while at the first and two-year follow-up waves, a subset of the HOME items in each wave is dedicated to estimate the stimulation factor, and a different subset of items is dedicated to estimate the warmth factor. Importantly, we exclude from this estimation any items directly related to the intervention (e.g., presence of children's books, which were provided as part of the intervention), as well as items with low factor loadings (below 0.4).

Finally, using the same methodology, we specify fully dedicated factor models at baseline and the two follow-up waves to estimate two additional parental latent factors, separately for each survey wave: a knowledge factor that uses individual items of the parental knowledge of child development scale, and a wellbeing factor that uses as inputs the raw scores of different scales (such as the CES-D scale to measure depressive symptoms) collected in each wave.[12]

4.2 Main Estimation strategy

We use our estimated factors at each wave and the randomized nature of the *Msingi Bora* intervention to estimate the causal impact of the program on our child and parental outcomes of interest in both the short-term and two years later. Short-term impacts are estimated with the following model using data from the first follow-up survey:

$$Y_{iv1} = \alpha_0 + \gamma_1 Groups_{i,v} + \gamma_2 Mixed_{i,v} + X_{iv0}'\delta + \varepsilon_{iv1} \qquad (2)$$

where $Y_{iv1} = \{\theta_{iv1}^{k}, \theta_{iv1}^{l}\}$ denote the error-free child factor k or parental factor l, for household i in village v, measured at the first follow-up survey (t=1). The coefficients of interest are γ_1 and γ_2, which capture the intention-to-treat (ITT) impacts of the group-only and mixed-delivery interventions, respectively. In the most parsimonious form of equation (2), the variable X_{iv0} controls only for randomization strata (the subcounty) and the baseline outcome, while in richer versions, to increase precision, it additionally controls for other baseline characteristics as prespecified in an analysis plan and that include an asset index measure of household wealth, maternal education, child sex, and birth order. Finally, ε_{iv1} is a disturbance term clustered at the level of villages *v,* the unit of randomization.

[12] As described in Section 3.4, wellbeing measures at baseline include depression and social support; at the first follow-up wave they include depression, stress, social support, self-efficacy, and beliefs about the importance of parenting; and at the two-year follow-up survey they include depression, social support and beliefs about the importance of parenting.

We estimate the sustained effects of *Msingi Bora* after two years, as well as the differential impacts of booster villages, based on regressions of the following form using data from the second follow-up survey:

$$Y_{i,v,2} = \alpha_0 + \gamma_1(Groups * No\ Boosters)_{iv} + \gamma_2(Groups * Boosters)_{iv} + \gamma_3(Mixed * No\ Boosters)_{iv} + \gamma_4(Mixed * Boosters)_{iv} + X'_{iv0}\delta + \varepsilon_{iv2} \quad (3)$$

where $Y_{iv2} = \{\theta^k_{iv2}, \theta^l_{iv2}\}$ is the outcome vector comprising child and parental factors at the two-year follow-up survey ($t = 2$). Here, variables are defined as in equation (2), but coefficients of interest now differentiate between sustained impacts of the original intensive phase of *Msingi Bora* two years after the end of the program (γ_1 and γ_3), and the combined effect of the original intensive intervention with the additional boosters between follow-up survey waves (γ_2 and γ_4). We internally standardize all estimated child and parent latent factors in a given wave using 2-month age bands relative to the control group means, estimated nonparametrically to have zero mean and unit variance. Though our creation of latent factors significantly reduces the number of individual hypotheses to test, our setting with multiple treatment arms and outcomes necessitates we use the Romano-Wolf correction to address multiple hypothesis testing (Romano and Wolf 2005).

4.3 Heterogeneous Treatment Effects

We examine heterogeneity in treatment effects to understand who benefited most from participation in *Msingi Bora* and examine if the program helped close any socioeconomic gaps using the Sorted Partial Effects method (SPE) recently developed by Chernozhukov, Fernández-Val, and Luo (2018). Using SPE allows us to consider all sources of heterogeneity at once rather than considering one characteristic at a time to have a richer portrait of heterogeneity in impacts. Moreover, these sorted effects form the basis for classification analysis (CA), which examines how the most and least affected children and families differ along a wide range of observed characteristics.

To capture full heterogeneity of treatment effects, we consider an interactive linear model where treatment assignment is allowed to interact with all relevant covariates in the following equation, where for simplicity we suppress the village subscript:

$$Y_{i,2} = Z'_{i,h}\theta + \varepsilon_{i,h,2} \quad (4)$$

Where $Z_{i,h} = (D_{i,h}, Q_{i,h})$ is a vector containing dummies for the 4 treatment assignments, which for simplicity of notation we denote by $D_{i,h}$ (e.g., $h = 1\ for\ Groups\ and\ No\ Boosters$), as well as the vector $Q_{i,h}$, which includes the same set of covariates $X_{i,0}$ as in equation (3) as well as their interactions with treatment dummies, $X_{i,0} * D_{i,h}$, to fully capture heterogeneous impacts by individual characteristics. The predictive partial effect (PE) of a given treatment h relative to the control group, and for a specific value of the vector Q, denoted by q, is:

$$\tau_h(q) = (1, q)'\theta - (0, q)'\theta \quad (5)$$

Chernozhukov et al. (2018) propose to report the entire set of PEs sorted in increasing order and indexed by percentiles $u \in [0, 1]$ with respect to the distribution of the covariates in the population of interest. The uth percentile of $\tau(Q)$ is then the uth-Sorted Partial Effect (u-SPE). The sample analog of τ, $\widehat{\tau(q)}$, is obtained by just replacing the estimated θ in equation (5) by the OLS estimates $\hat{\theta}$ obtained from equation (4). Displaying the SPE by increasing values of u enables us to visualize the full range of the heterogeneous effects in one single plot.

Finally, we use our estimated SPEs to carry out classification analysis, which consists of classifying children's and parental outcomes into most or least affected by intervention h depending on whether their PEs are above or below some tail SPE, and then compare how the most and least affected groups differ in observable characteristics. In section 6, we report differences between the upper and lower halves of the effect distribution.

4.4 Mechanisms and Mediation Analysis

A key hypothesis of our experiment was that *Msingi Bora* would improve children's outcomes by changing parenting behaviors and other inputs such as parental knowledge and wellbeing. To explore these hypothesized mechanisms, we conduct a standard mediation analysis to explore to what extent concurrent intervention impacts in parental inputs at the two-year follow-up (FU2), as well as past impacts on children's outcomes at the first follow-up (FU1), can account for observed impacts on child outcomes at the time of the two-year follow-up survey.

Building on our main empirical model in equation (3), we estimate the augmented equation:

$$\theta_{i2}^{k} = \alpha_0 + \textstyle\sum_{h=1}^{4} \beta_h D_{ih} + X_{i0}' \delta_5 + \sum_l \rho_6^l \, \theta_{i2}^{l} + \sigma_7^k \theta_{i1}^{k} + w_{i2}, \qquad (6)$$

where X_{i0} includes baseline measures of child skill k and is as defined in equation (2). Using this model, the goal is to examine how the estimated β_h coefficients in equation (6) change relative to the estimated ITT γ_h coefficients from equation (3), when we now include current parental outcomes θ_{i2}^{l}, and past child outcomes, θ_{i1}^{k}, as potential mediators.[13]

Note that this mediation analysis is necessarily exploratory and non-causal because the mediators included in equation (6) are potentially endogenous for two reasons. First, there may be other unobserved parental inputs that change with the intervention and correlate with our observed mediators, and second, there is potential reverse causation from child outcomes to mediators. However, this analysis is still valuable to determine the potential strength of the hypothesized mechanisms relating parental behaviors and child outcomes over time to explain our impacts at the second follow-up survey. Note also that we include past changes in child outcomes induced by our program rather than past changes in parental inputs as a mediating pathway under the assumption that child outcomes more directly capture the cumulative history of parental inputs between baseline and the first follow-up survey.

---

[13] The β_h in equation (6) can be interpreted as the program's "direct effects," and capture the effect of unmeasured inputs.

The final mediation model includes all mediators that are found to be relevant, that is, those that are a) statistically affected by the intervention in equations (2) or (3), and b) statistically significant when included alone in a version of equation (6). We estimate the model in steps using a Monte Carlo simulation approach following Campos et al. (2017). For example, focusing on parental outcomes at the second follow-up survey, we first obtain the regression coefficients γ_h from equation (3). Second, we obtain estimates of ρ_6^l from equation (6). Under the very strong assumption of orthogonality between $\theta_{i,2}^l$ and $w_{i,2}$, we can interpret the product $\gamma_h * \rho_6^l$ as the component of the overall ITT impact in arm h coming through parental outcome $\theta_{i,2}^l$, which is denominated as the intervention's "indirect effect." An equivalent analysis could be performed to examine past child outcomes as mediators.[14]

Next, we compute the 95% Monte Carlo confidence intervals for the indirect effect of each mediator based on 20,000 repetitions. A confidence interval that does not include zero indicates a significant indirect effect of that particular mediating variable on child outcomes. Finally, for the total indirect effect, we include all the relevant mediators in the model at once.

5 Main Results

5.1 Short-Term Impacts

Table 3 summarizes our updated estimates of the short-term ITT impacts of *Msingi Bora's* first intensive phase of 16 sessions using our latent indices of child and parental outcomes. Results are very consistent with those based on the individual measures (Luoto et al 2020). At the first follow-up survey, children from villages that received only group meetings have 0.49 SD larger cognitive scores than children from control villages, while children from mixed-delivery villages have 0.27 SD larger cognitive scores (column 1). For parental outcomes, both delivery models result in large effects on our index of maternal stimulation of roughly 0.8 SD (column 2), and smaller effects on the maternal knowledge index (column 5), but neither model significantly improves the maternal warmth index (column 3). The group-only delivery model has a marginally significant 8 percentage point increase in the share of mothers reporting feeding their children at least 4 of 7 food groups in the past 24 hours, while the mixed-delivery model has no impact on nutrition behaviors (column 4). Comparing the two delivery models, differences in estimated impacts are not statistically significant under a two-sided test for any primary outcome, but group-only delivery has larger impacts based on one-sided tests for children's cognition (p=0.085) and nutrition behaviors (p=0.057). Neither model of delivery significantly improves our index of maternal wellbeing (column 6).

Table 3: Short-Term ITT Impacts on Child and Parent Factors

	(1) Cognitive	(2) Stimulation	(3) Warmth	(4) Nutrition	(5) Knowledge	(6) Wellbeing

[14] A new first stage equation analogous to equation (6) should characterize the ITT impacts on children's outcomes measured at FU1 to extend this analysis.

	Index	Index	Index	(0/1)	Index	Index
1. Groups Only	0.487***†	0.814***†	0.086	0.079*$^{+}$	0.196***†	-0.063
	(0.173)	(0.122)	(0.147)	(0.044)	(0.066)	(0.108)
2. Mixed-delivery	0.267*$^{+}$	0.817***†	0.097	0.013	0.135**†	-0.150
	(0.159)	(0.117)	(0.121)	(0.048)	(0.067)	(0.097)
Observations	1070	1070	1070	1063	1070	1063
p-value for Mixed>Groups	0.085	0.513	0.529	0.057	0.149	0.191

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses are clustered at the village level. Each column is a separate regression for each outcome. Parental outcomes in columns 2-5.$^{+}$Signifies estimate is significant at 10% level using Romano-Wolf estimator that corrects for multiple hypothesis testing. †Signifies estimate significant at 5% level using Romano-Wolf estimator. Data are in SD. Effect sizes and p-values obtained from intention-to-treat (ITT) estimates for each outcome measured at first follow-up survey using internal age-standardization to the control group. All results include adjustments prespecified in our study protocol: child's age, household wealth, maternal education, child sex, birth order, the outcome at baseline (if measured) and sub-county fixed effects (the strata). P-values at bottom of table are based on one-sided Wald tests of the null hypothesis that mixed-delivery would outperform group-only delivery.

5.2 Two-year Impacts

Children's development

Table 4 presents the two-year ITT impacts of *Msingi Bora* on our indices of children's cognitive (column 1) and socioemotional (column 2) development at the two-year follow-up survey. Children in villages that received only group sessions during the original intensive phase of *Msingi Bora*, without subsequent boosters, show 0.20 SD higher sustained scores two years later for both cognitive and socioemotional outcomes than do children in the control group. The combination of the group-only intervention followed by subsequent boosters leads to 0.23 SD and 0.24 SD higher gains in children's cognitive and socioemotional scores, respectively, relative to children in the control group.[15]

For the mixed-delivery model, without boosters, we find no evidence of sustained impacts on children's cognitive or socioemotional scores after two years and point estimates are near zero. Meanwhile, combining the mixed-delivery model with subsequent boosters leads to 0.24 SD higher socioemotional scores as compared to children in the control group, but the estimated 0.11 SD impact on children's cognitive scores is not statistically significant.

In total, the booster extension appears to help improve children's socioemotional but not cognitive development: one-sided Wald tests of the combined impacts from all booster villages being greater than those from all non-booster villages has a p-value of 0.05 for the socioemotional index versus 0.27 for the cognitive index.

Finally, though we find no evidence of differential attrition in section 3.4, Table A4 in the Appendix shows that all our impacts on child outcomes are robust to potential selective attrition corrected using Lee bounds (Lee 2009).

[15] For reference, Table A3 in the Appendix presents results for the same specification but using individual measures children's cognitive and socioemotional development as outcomes.

Table 4: Two-year ITT Impacts on Child Cognitive & Socioemotional Factors

	(1) Cognitive Index	(2) Socioemotional Index
1. Groups No Boosters	0.200**+	0.204**†
	(0.099)	(0.082)
2. Groups + Boosters	0.230**†	0.237***†
	(0.089)	(0.085)
3. Mixed No Boosters	0.044	0.036
	(0.108)	(0.090)
4. Mixed + Boosters	0.110	0.237**†
	(0.100)	(0.091)
Observations	944	944
P-value Boosters>Non-Boosters	0.27	0.05

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses are clustered at the village level. +Signifies estimate is significant at 10% level using Romano-Wolf estimator that corrects for multiple hypothesis testing. †Signifies estimate significant at 5% level using Romano-Wolf estimator. Data are in SD. Effect sizes and p-values obtained from intention-to-treat (ITT) estimates for each outcome measured at the two-year follow-up using internal age-standardization to the comparison group. All results include adjustments prespecified in our study protocol: child's age, household wealth, maternal education, child sex, birth order, the outcome at baseline (if measured) and sub-county fixed effects (the strata). P-values at bottom of table are based on one-sided Wald tests of booster villages having greater impacts than non-booster intervention villages.

Parental behaviors, knowledge and well-being

Table 5 presents results on the two-year impacts of *Msingi Bora* on our indices of maternal stimulation (column 1), warmth (column 2), and nutrition behaviors (column 3), as well as knowledge (column 4), and wellbeing (column 5). Mothers from villages that received only group sessions, without subsequent boosters, sustain 0.19 SD (stimulation) and 0.34 SD (knowledge) higher scores than mothers in control villages. There is no evidence of sustained impacts on the other maternal outcomes for this arm. Villages that received a combination of group sessions with the booster extension have statistically significant impacts on maternal stimulation (0.27 SD), knowledge (0.37 SD), and nutrition (9 percentage point improvement), but there are no statistically significant impacts on maternal warmth or wellbeing for this arm.

For those villages that initially received the mixed-delivery model and no booster extension, all impacts after two years fade from statistical significance after correcting for multiple hypothesis testing. The combination of a mixed-delivery model with subsequent boosters leads to improvements of 0.39 SD and 0.25 SD in maternal stimulation and knowledge of child development, respectively, but no impacts are observed on any other outcome. Table A5 in the Appendix shows that all statistically significant impacts on parental outcomes are robust to potential selective attrition corrected using Lee bounds.

Overall, we see suggestive evidence that boosters helped to improve maternal stimulation behaviors (p-value 0.07 on the one-sided test in the bottom row of Table 5). Results for all other parental outcomes in Table 5 show that point estimates are consistently larger for villages that received the booster extension, though not always by statistically significant amounts.

Table 5: ITT Impacts on Parental Behavioral Indices at Follow-Up 2

	(1) Stimulation Index	(2) Warmth Index	(3) Nutrition (0/1)	(4) Knowledge Index	(5) Wellbeing Index
1. Groups No Boosters	0.190*+	-0.046	0.043	0.338***†	-0.116
	(0.097)	(0.081)	(0.042)	(0.078)	(0.124)
2. Groups + Boosters	0.266**†	0.140	0.093**+	0.361***†	0.061
	(0.130)	(0.108)	(0.044)	(0.111)	(0.122)
3. Mixed No Boosters	0.214*	0.123	-0.028	0.157	-0.073
	(0.117)	(0.119)	(0.040)	(0.123)	(0.096)
4. Mixed + Boosters	0.392***†	0.164	-0.008	0.250**+	0.066
	(0.113)	(0.109)	(0.045)	(0.121)	(0.121)
Observations	942	942	941	944	939
P-value Boosters > Non-Boosters	0.073	0.101	0.150	0.272	0.056

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses are clustered at the village level. +Signifies estimate is significant at 10% level using Romano-Wolf estimator that corrects for multiple hypothesis testing. †Signifies estimate significant at 5% level using Romano-Wolf estimator. Data are in SD. Effect sizes and p-values obtained from intention-to-treat (ITT) estimates for each outcome measured at the two-year follow-up survey using internal age-standardization to the comparison group. All estimations presented include controls prespecified in our study protocol: child's age, household wealth, maternal education, child sex, birth order, the outcome at baseline (if measured) and sub-county fixed effects (the strata). P-values at bottom of table are based on one-sided Wald tests of all booster villages having greater impacts than all non-booster intervention villages.

5.3 Comparing short-term vs. medium-term impacts

With or without boosters, the impacts after two years on children's cognition of roughly 0.20 SD in Table 4 under the group-only delivery model are noticeably smaller than the short-term impacts of 0.49 SD in Table 3. Impacts fade entirely from significance for the mixed-delivery arm, and similarly smaller impacts can be seen for parental stimulation under both delivery models comparing Tables 5 and 3. To understand why impacts decline over time, we do the following. Instead of age-standardizing our latent child cognitive factors relative to the control group within a given wave, we re-standardize them relative to the control group at baseline, which we anchor at zero. This allows us to compare the trajectories of cognitive development across survey waves, by treatment arm. For expositional purposes, we focus on the group-only arm, which has detectable sustained impacts after two years, and we pool booster and non-booster villages to improve our power since boosters had no added cognitive impacts.

Figure 3 shows that, between baseline and the first follow-up survey roughly a year later, children in the group-only arm experienced an increase in their cognitive scores of about 0.23 SD relative to the mean scores of the control group at baseline. Simultaneously, children in the control group experienced a decline in cognition of about 0.26 SD relative to their baseline scores.[16] These combined effects account for the total estimated short-term impacts on cognition for this arm reported in Table 3. The observed decline in age-standardized cognitive scores for children in the control group between baseline, when children were 14 months on average, and the first follow-up survey, when children were 26 months on average, is consistent with evidence of cumulative deficits starting to emerge around age 1-2 years for children in many LMIC settings (Black et al. 2017; Engle et al. 2011; Fernald et al. 2006).[17]

In the two years between follow-up surveys, we see a slight decline in the mean standardized cognitive scores for the group-only arm, though they remain statistically positive. Meanwhile, the mean cognitive scores for children in the control group slightly improve, and this partial "catch up" explains why net treatment impacts decline from 0.49 SD at the first follow-up survey to 0.21 SD at the two-year follow-up survey. We hypothesize that the slight decline in cognitive scores among treated children between follow-up surveys can be partially explained by parents not adapting their parenting practices as children get older and/or not sustaining the learned new behavior over time.[18] Meanwhile, the "catch-up" among untreated children is consistent with children engaging more with their environments, entering preschool, and generally becoming less dependent on parents for stimulation as they get older. Figure A2 in the Appendix shows that this "catch-up" effect among children in the control group is almost entirely driven by children from relatively better-off families, for whom preschool attendance is substantially higher. For poorer children in the control group, we see no evidence of any "catch up."

Figure 3: Trajectories of children's cognition by group-only vs control arms

[16] We anchor cognitive scores of children in the control group at baseline to zero, but Figure 1 shows results after adjusting for prespecified baseline characteristics including maternal age and education, a wealth index, children's sex and birth order, and subcounty strata fixed effects. These adjustments plot the mean outcome for this subgroup slightly off from zero in the Figure.

[17] We find an identical pattern to Figure 1 if we instead use just the individual measures of child cognition from each wave and perform a similar analysis using the official age-standardizations of scores that are based on international (US) norms for children's development. We present results for our analysis of trajectories using our latent cognitive factors that encompass more than a single measure in each wave.

[18] Figure A1 in the appendix shows that stimulation scores also decrease between follow-up surveys among treated parents, a result also shown across Tables 3 and 4.

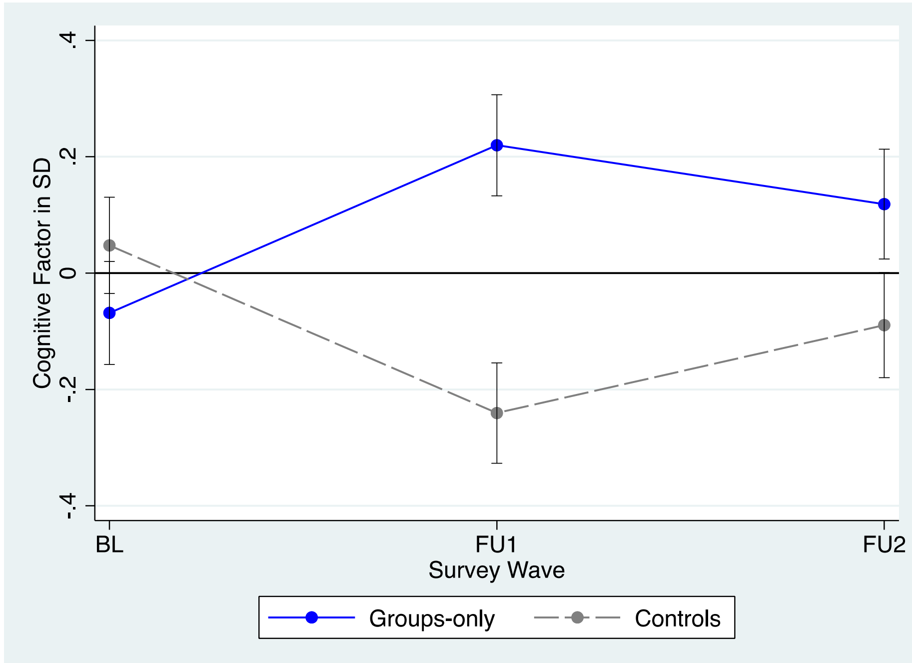


Figure plots results for mean age-standardized cognitive factors along with 90% confidence bands for group-only and control group children at each survey wave. Each arm at each wave is age-standardized relative to the baseline control group, anchored at zero. All estimates adjust for prespecified baseline characteristics including maternal age and education, a wealth index, children's sex and birth order, and subcounty strata fixed effects.

6 Additional Results

6.1 Heterogeneous Treatment Effects

A key question for policy makers is whether parenting programs can help address early developmental disadvantages among the poorest children and whether the effects are sustained into later childhood. Studies from high-income countries have found larger sustained gains in children's cognitive outcomes for lower-educated households (Carneiro et al. 2023; Doyle 2020) and in socioemotional outcomes for girls versus boys (Heckman et al. 2017). To our knowledge, only one study of sustained outcomes from a LMIC setting examined heterogenous effects and found no differential impacts across any child or household characteristic two years after the end of their program (Justino et al. 2022).

To examine heterogeneity in impacts after two years and from our additional boosters, we first estimate sorted partial effects (SPE) as in Chernozhukov et al., (2018) and then we conduct a classification analysis to compare sociodemographic characteristics across those "most" versus "least" affected by the intervention, defined as those below or above the median predicted SPE, respectively. Table 6 presents heterogeneous impacts for our main outcomes of interest, children's cognitive and socioemotional development, as well as parental stimulation behaviors.

The clearest finding that emerges from Table 6 is that impacts on children's cognition and parental stimulation behaviors are largest among relatively poorer households, with or without the booster extension. For example, across all villages in the group-only arm (panel 3 in Table 6), children who benefited the most from the intervention in terms of their cognitive development come from households that are 1.6 SD poorer on average (using our standardized wealth index) than households with children who benefited the least. We see a similar pattern for parental stimulation behaviors, and with or without boosters. Though no households in our sample would truly qualify as objectively "wealthy" this result is encouraging if it suggests our program was protective against fade-out among the most vulnerable homes for these outcomes. A comparison of the predicted SPEs along the entire distribution of wealth for our sample shows that these estimated differences are not just at the median, but are true for the entire distribution of SPEs (Figure 4).

We do not observe corresponding differences by wealth on children's socioemotional outcomes in Table 6. In fact, we see that children with the largest impacts on socioemotional outcomes have mothers with more education, particularly in villages that did not receive boosters. Since our socioemotional index is highly driven by children's prosocial behaviors and conduct problems, this result suggests that more educated mothers were most able to adopt positive parenting practices to manage child behaviors. We also see larger socioemotional impacts on boys, a result that is somewhat contrary to the socioemotional impacts reported in the studies mentioned above that either find no differential impacts by gender or positive impacts only for girls, 3 to 6 years after the end of their parenting programs.

Finally, we observe stronger effects on parental stimulation behaviors for households where the target child was second or higher birth order, compared to firstborns. This result is driven by booster villages, though it does not translate into differential gains by birth order in children's outcomes.

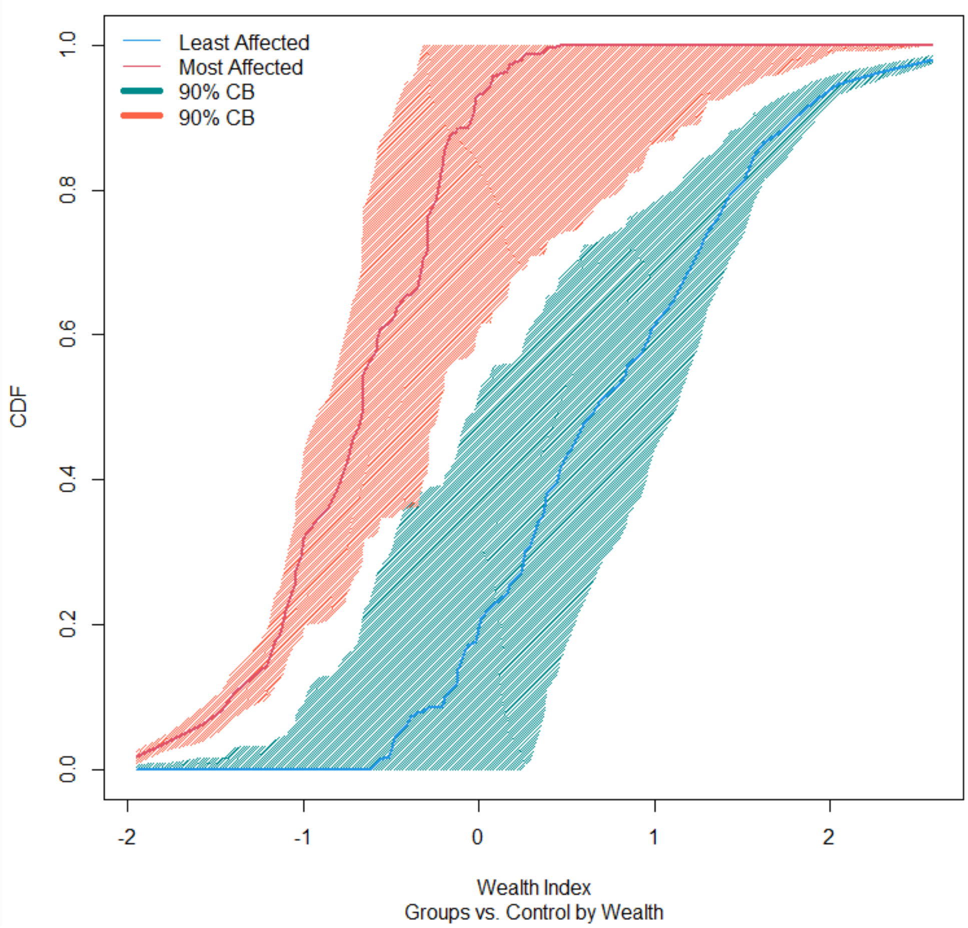
Table 6: Classification Analysis: Most vs. Least Affected Subgroups

	Cognition	**Socioemotional**	**Stimulation**
1) Groups + Boosters			
Mother Age	3.08	-3.35	-1.16
Wealth Index	-0.93***/+	-0.61	-0.85***/+
Mother Education (years)	-1.24*	1.89	-0.96*
Child Baseline index	0.64**	-0.37	-0.09
First Born (0/1)	-0.02	-0.11	-0.40***/+++
Girls (0/1)	-0.28	-0.81***/+++	-0.07
2) Groups + Non Boosters			
Mother Age	-9.64***/+++	-2.40	-6.28***/+
Wealth Index	-1.39***/+++	0.09	-1.06***/++
Mother Education (years)	0.65	3.69***/+++	1.55**
Child Baseline index	-0.69*	0.10	0.09
First Born (0/1)	0.19*	-0.03	-0.24*
Girls (0/1)	-0.03	-0.61***	0.18
3) Groups Overall			

Mother Age	-5.50**	-3.43	-3.92**
Wealth Index	-1.59***/+++	-0.24	-1.07***/++
Mother Education (years)	-0.16	3.29***/+++	0.13
Child Baseline index	-0.29	-0.28	-0.17*
First Born (0/1)	0.08	-0.06	-0.41***/++
Girls (0/1)	-0.29	-0.74***/++	0.01

Notes: Results based on comparisons above/below medians for sorted partial effects, where differences are calculated as the mean characteristic of the most affected minus mean characteristics of the least affected, split at the median. *p<0.10, **p<0.05, ***p<0.01 are p-values for statistical significance based on single hypothesis tests. +p<0.10, ++p<0.05, +++p<0.01 are p-values for joint statistical significance that account for multiple hypothesis testing.

Figure 4: Sorted Partial Effects of Cognition after Two Years: Most vs. Least Affected, by Wealth Index



Notes: Figure plots predicted SPEs at the two-year follow-up survey over baseline wealth index and associated 90% confidence bands using 500 bootstrap replications for those “most affected” (above the median SPE) and “least affected” (below the median SPE) by the group-only intervention at the two year follow-up survey.

6.2 Mediation Analysis

Our results above show that families who received the group-only *Msingi Bora* intervention, both with and without subsequent booster sessions, demonstrate sustained improvements in children's cognitive and socioemotional development, as well as in parental stimulation and knowledge of child development at the two-year follow-up survey.

In this section, we estimate a standard mediation analysis model to shed light on the potential mechanisms underlying these sustained impacts. In particular, we examine to what extent the two-year impacts for children's cognitive and socioemotional scores can be explained by concurrent impacts on parental stimulation behaviors and knowledge induced by the intervention, as well as past impacts in children's cognition as measured at the first follow-up survey. As is usual in this type of analysis, the assumptions required to decompose medium-term (two-year) treatment effects into direct and indirect effects are strong, and our mediation results can only be interpreted as suggestive evidence of the importance of these channels.

Table 7 presents results from estimating equation (6) for our cognitive and socioemotional factors at the two-year follow-up survey (FU2). As points of reference, columns 1 and 3 report the ITT coefficients of the two-year impacts of the program on children's cognitive and socioemotional outcomes (repeated from Table 4). Columns 2 and 4 include all relevant mediators, defined as those that are changed by the intervention, as well as have a statistically significant individual association with the corresponding child outcome at the two year follow-up survey.[19] In both cases, the pool of relevant mediators includes past experimentally-induced impacts on children's cognition measured at the first follow-up survey (FU1), as well as concurrent impacts on maternal stimulation and knowledge (nutrition is not a relevant mediator for any child outcome and is therefore dropped from analysis).

When all significant mediators are added to the model in column 2, the total estimated impact on children's cognition declines from 0.23 SD to 0.02 SD for villages that received group-only sessions and the booster extension. This means that combined, these three relevant mediators can explain up to 93% of the overall impact on children's cognition for this arm, a large indirect effect. In group-only villages that did not receive boosters, the total ITT cognitive impact of 0.20 SD in column 1 declines to 0.04 SD in column 2 when all relevant mediators are included, which translates into an indirect effect of 78%. Slightly smaller mediating effects are seen for socioemotional outcomes in column 4. The overall impact on socioemotional development of 0.24 SD for booster villages declines to 0.06 SD when relevant mediators are jointly included in the regression model, which implies an indirect effect of 72%. In non-booster villages, the total impact of 0.20 SD declines to 0.08 SD when we include all relevant mediators jointly, and the indirect effect is 60%.

Next, we decompose the relative importance of each relevant mediator to see how much each contributes to the total estimated program impacts on child cognitive and socioemotional outcomes at the two-year follow-up survey. Figure 4 shows that for children's cognitive

[19] Results for each individual mediator are presented in Tables A5 and A6 in the Appendix for cognitive and socioemotional development, respectively, and show the same pool of relevant mediators are included in columns 2 and 4 of Table 7.

outcomes, the most important mediator is past impacts on children's cognition at FU1, which explains up to 44% of the overall impact on present cognition in booster villages and 40% in non-booster villages. Concurrent two-year impacts on maternal stimulation is the second most important mediator, explaining up to 30% of the two-year impacts on cognition in booster villages and up to 21% in non-booster villages. Finally, concurrent changes in maternal knowledge of child development at FU2 explain up to 19% of the two-year impacts on cognition in booster villages, and up to 17% in non-booster villages.

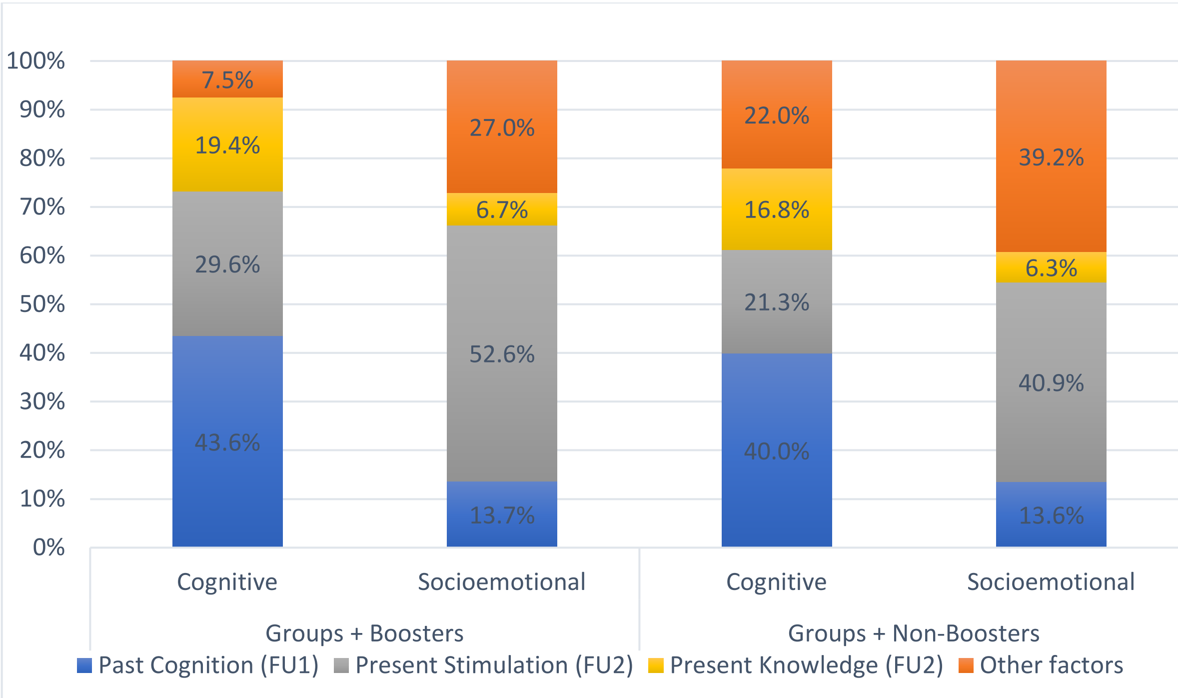
For socioemotional outcomes, we observe that the mediating power of past impacts on children's cognition is significantly lower after two years, explaining only 14% of the effect of the program both in booster and non-booster villages. Concurrent two-year improvements in maternal stimulation is by far the most important mediator for socioemotional outcomes, explaining up to 53% of the sustained impacts on socioemotional scores in booster villages and up to 41% in non-booster villages. Present two-year impacts on maternal knowledge have a small mediating role for socioemotional outcomes, explaining just 7% and 6% for both booster and non-booster villages, respectively.

Table 7: Mediation Effects on Children's Cognitive and Socioemotional Development at Follow-up 2

	Cognitive Index		Socioemotional Index	
	(1)	(2)	(3)	(4)
Groups + Boosters	0.228**	0.017	0.233***	0.063
	(0.088)	(0.081)	(0.084)	(0.065)
Groups + Non-boosters	0.200**	0.044	0.204**	0.080
	(0.099)	(0.096)	(0.082)	(0.097)
Changes in Children's Cognition at FU1		0.174***		0.058**
		(0.031)		(0.028)
Changes in Stimulation at FU2		0.235***		0.444***
		(0.039)		(0.031)
Changes in Knowledge at FU2		0.111***		0.041
		(0.034)		(0.034)
Observations	945	915	945	915
% Indirect Effect Groups+Boosters		**92.5%**		**72.4%**
% Indirect Effect Groups+No Boosters		**78.0%**		**60.0%**

Notes: Results from estimation of equation 7. Table excludes mixed-delivery arm. Full mediation results for each individual mediator are in Appendix Tables A6 and A7 for cognitive and socioemotional development, respectively.

Figure 4: Mediation Analysis Children's Outcomes at Follow-up 2



Notes: Figure shows the share of a given outcome at the two-year follow-up survey "explained" by the relevant mediator for an intervention group of interest. Results based on Table 7.

6.3 Benefit-Cost Analysis

A cost-effectiveness analysis previously undertaken based on the short-term results found that *Msingi Bora's* group-only model was the most cost-effective intervention to date in a LMIC setting (Lopez Garcia, Saya, and Luoto 2021). We estimated a total direct implementation cost per child of $119 for the 8-month intervention using only group sessions. We also found the program could be much more cost-efficient if CHVs delivered more than one group session every two weeks, so we view these costs as an upper bound on any scaled version of the program. The highest cost categories were travel and full-board accommodation for the two centralized trainings in Kisumu. However, simply by switching to a training model that takes place within subcounties, which we successfully adopted during the booster extension even before the onset of the COVID-19 pandemic, we estimate a scaled version of *Msingi Bora* could cost $90 per child, a 24% savings overall.

An accompanying benefit-cost analysis (BCA) based on the short-term results that incorporated a societal perspective and predictions about future benefits and costs estimated a benefit-cost ratio (BCR) of 15.5 for the group-only delivery model (Lopez Garcia et al. 2021). Societal costs in this estimate account for mothers' opportunity costs to attend the group sessions and to enact the behavioral changes over the subsequent three years. Future benefits include expected gains in lifetime wages, and future costs include increased schooling costs, both stemming from higher child cognitive skills imparted by the program. An accompanying sensitivity analysis calculates that program benefits would remain greater than costs for any sustained cognitive impact of at least 0.03 SD.

Here, we now update our cost estimates and associated BCRs based on findings from the two-year follow-up survey including the emergence of separate impacts on children's socioemotional skills, and with the inclusion of the extended booster sessions and their associated costs. As before, we adopt a societal perspective that accounts for program implementation costs as well as present and future societal costs and benefits.

Table 8 summarizes the additional costs of delivering the extended booster sessions over the timeframe November 2019 to July 2021 (column 2), as well as reprints the calculated costs from delivery of the group-only model during the first intensive phase of *Msingi Bora* as published in Lopez Garcia et al. (2021) (column 1). Panel A summarizes the direct implementation costs for things like labor and training, which we use for comparison with other parenting interventions from LMICs. Panel B in Table 8 additionally accounts for societal costs in the form of the opportunity costs of mothers' time to attend the booster sessions and enact the behavioral changes, where mothers' time was priced at local wage rates. To enable direct comparisons with our published short-term results, we convert all prices to January 2020 US dollars.

In terms of direct implementation costs, labor costs account for CHV time to attend trainings, mobilize families in their village, as well as travel to and host the sessions. Relative to the main intervention, this cost increased as CHVs hosted the same booster session more than once per village to accommodate smaller groups following the onset of the COVID-19 pandemic. All other costs decreased relative to the main intervention including the time costs of supervisory personnel to oversee a total of 9 booster sessions over 22 months per village versus a more intensive schedule of 16 group sessions over 8 months per village during the original program. Training costs for the boosters were also significantly reduced by being conducted locally within subcounties versus centralized trainings that necessitated full-board accommodations. Other savings from the booster extension relative to the original trial included mobile airtime for CHVs and supervisors, and we did not reprint t-shirts or certificates of completion for the booster extension. The overall cost per child for the booster extension is $37.61in 2020 US dollar terms, or 32% of the original cost of the 8-month group-only program.

With regards to societal costs, the original estimate for mothers' opportunity costs of their time in column 1 for the original trial included time to attend sessions, as well as to enact the practices over a three-year timeframe. We therefore do not add further opportunity costs for maternal behavior change during the booster extension and only account for their time to attend the additional sessions. The overall societal cost per child for the booster extension is $2.85 in 2020 US dollar terms. More details on the unit costs for these estimates can be found in Lopez Garcia et al. (2021).

Table 8: Program Implementation Costs: Original *Msingi Bora* + Extended Boosters

	(1)	(2)
	Group-only arm	**Booster Sessions**
Panel A: Direct implementation costs		
Personnel – CHVs	$3,570	$4,072
Personnel – Supervisory Staff	$16,674	$4,515

Training & Travel Costs	$22,449	$4,855
Other Costs (printing manuals, airtime supervisors, start-up costs)	$4,730	$1,601
Direct implementation costs per child	**$119**	**$38**
Panel B: Societal costs	**$8,748**	**$1,142**
Mothers' opportunity costs	$7,096	$1,142
Venue costs	$1,652	$0
Societal costs per child	**$22**	**$3**

Notes: All costs originally in Kenyan Shillings converted to USD using exchange rate of 1 USD = KSh 101.8 as of January 2020. In practice, venue costs were always free, but a reviewer of our published short-term results argued (successfully) that we cost out how much venue rentals would be in a scaled version of *Msingi Bora*, which are included in the table in column 1. For the booster extension, since CHVs hosted smaller booster sessions outside following the outbreak of COVID-19, we do not again include venue costs.

The costs summarized in Table 8 include only present period costs that ignore the future streams of potential costs and benefits that may result from the program and are necessary to calculate long-term benefit-cost ratios. Future costs include additional schooling costs associated with increased school attendance stemming from the program. Future benefits include expected returns in lifetime wages due to increases in both cognitive and now newly emerging socioemotional skills. Because our sample is only aged 3.5-5 years, we need to translate our two-year program impacts on children's cognitive and socioemotional development into impacts on adult outcomes (school attendance and wages). We do so by following the same strategy we employed for our short-term analysis, the details of which we explain in Appendix B.

To maintain comparability with our earlier estimated BCRs based on the short-term results, we use a conservative discount rate of 5% to express our predicted future costs and benefits in present discounted value (PDV) terms, though we test the sensitivity of our results to different discount rates. Also, this analysis is based on the strong assumption that our two-year medium-term impacts on cognitive and socioemotional outcomes are good predictions of long-term impacts on adult outcomes of schooling and wages. The fading impacts on child cognition already present after two years suggests this may be risky. We therefore calculate what size impacts must remain to preserve positive benefit-cost ratios including the additional costs of the boosters.

Table 9 summarizes results for the group-only arm, both with and without the booster extension. Comparing columns 1 and 2, the total ITT program impacts on wages is slightly higher with the boosters' slightly larger impacts on both cognition and socioemotional outcomes, resulting in slightly higher gains in lifetime earnings per child of $195 in 2020 prices. However, with the additional costs of the booster extension combined with the higher implied long-term schooling costs, the resulting BCR is slightly lower at 6.9 versus 7.5 without the supplemental boosters.

We caution against using these numbers as rationale against continuing program support in the form of booster sessions for a few reasons. First, the booster sessions, but not the original program, had to survive the outbreak of the COVID-19 global pandemic and all its myriad complications. Second, the relatively stronger impacts of boosters on parenting behaviors in Table 5, combined with the strong mediating power of parenting behaviors on children's cognitive and socioemotional outcomes in Table 7, suggests it would be dangerous to conclude boosters were not a worthy (marginal) investment given that they effectively extended an original 8-month program for an additional 22 months for what amounts to a 32% increase in costs.

Under both models, the results in Table 9 show that even after two years we find our program continues to have projected benefits that outweigh costs by large margins, and this is true with and without the booster extension. Though these benefit-cost ratios are smaller than those based on the short-term results, both estimates remain markedly higher than the few existing estimates found in earlier home visiting parenting programs from both high- and low-income country settings (Heckman et al. 2017; Araujo et al. 2021). Considering the potential savings from a scaled version of *Msingi Bora*, these results are very encouraging for the program. We calculate that we would need to increase our chosen discount rate to over 13.5% (13%) to have costs outweigh benefits for *Msingi Bora* without (with) the supplemental boosters. Similarly, cognitive returns can fade to zero for the program with or without boosters and *Msingi Bora's* benefits still outweigh their costs thanks to the emergence of impacts on socioemotional outcomes.

Table 9: Benefit-Cost Ratios of Group-only *Msingi Bora* with and without Boosters, Two-Year Impacts

Delivery Model	(1) Without Boosters	(2) With Boosters
Discounted Sum of Lifetime Earnings per child (US $)	$13,880	$13,880
Total impact on wages (cognitive + socioemotional)	0.092	0.106
Gains in lifetime earnings per child	**$1,283**	**$1,478**
(1) Provider Costs per child	*$119*	*$156*
(2) Present period Societal costs per child	*$22*	*$25*
(3) Long-term societal costs per child (schooling costs)	*$30*	*$35*
Total societal costs per child (1+2+3)	**$171**	**$216**
Benefit-cost ratio	**7.5**	**6.9**

Notes: Intervention impact on wages calculated as the product of ITT intervention impact estimates from Table 4 and wage returns of 0.397 and 0.064 to cognitive and socioemotional outcomes as described in Appendix B. Gains in lifetime earnings per child calculated as discounted sum of lifetime earnings multiplied by intervention impact on wages. All costs originally in Kenyan Shillings converted to USD using exchange rate of 1 USD = KSh 101.8 as of January 2020. Provider costs per child include direct implementation costs as listed in Table 8. Present-period societal costs per child include mothers' opportunity costs of time, and for the program with boosters in column 2 includes costs from original program plus boosters. Long-term societal costs per child includes the predicted future costs of schooling. Discounted sum of lifetime earnings adjusts for expected survival probabilities using age life tables from Kenya. Discount rate for age earning profiles is 5%.

7 Discussion

In this paper, we present evidence from the evaluation up to two years after the end of a group-based parenting intervention explicitly designed to be affordable and potentially scalable for low-income, rural and resource-poor settings. We also test the role of extending program support after the end of the main intervention to help sustain improvements in parent and child outcomes over time. Our results show that the original more intensive *Msingi Bora* program had large positive impacts on children's cognition and parental stimulation behaviors immediately after its end. Effects are generally larger for villages that received only group visits, and persisted two years later only among these villages, albeit are smaller in magnitude. This decline in treatment effects between follow-up surveys is explained by a combination of declining impacts among treated children, as well as partial "catch up" among children in the control group, particularly children from relatively wealthy households. There is no similar catch-up effect among children from poorer households. Early impacts on parental knowledge remain high two years later.

The addition of supplemental, infrequent and "light touch" booster sessions has a small yet positive added value in helping to sustain impacts two years later, particularly for parental stimulation behaviors and children's socioemotional development, and despite the interruption of COVID-19 to their delivery. Two-year impacts on children's outcomes are largely mediated by concurrent impacts on parental stimulation and knowledge, as well as past impacts on children's cognition. These findings highlight the importance of early improvements in cognitive abilities as well as parental inputs such as book sharing, two-way conversations, and knowledge about optimal caregiving practices to promote sustained gains in cognitive performance. Taken together, we interpret results of our mediation analysis as suggestive of the presence of dynamic complementarities in parental investments over time, which are rarely demonstrated empirically in the literature, particularly from an LMIC, and highlight the importance of early intervention (Cunha, Heckman, and Schennach 2010; Cunha and Heckman 2007; Attanasio et al. 2020). Finally, our group-only *Msingi Bora* intervention, with or without the booster extension, remains highly cost-effective two years later, with projected benefits that outweigh costs roughly by a factor of 7, despite the reduced impacts after two years and the severe disruptions from the COVID-19 pandemic.

These findings are important for three reasons. First, despite the growing evidence base for positive short-term impacts of group-based parenting interventions, there is still very limited evidence of sustained impacts two or more years after the end of parenting interventions using any delivery model, and a common problem of fading impacts over time affect most programs (Jeong et al 2021). Our study demonstrates the sustained effectiveness of a low-cost and group-based ECD parenting intervention two years after the end of the intervention in a low-resource, rural setting, which is rare in the literature. The closest study to ours from an LMIC setting is the *First Steps* parenting program from rural Rwanda (Justino et al. 2022), which utilizes a radio show to communicate program messages to groups of parents, and finds that early impacts of 0.38 SD decline to 0.2 SD two years later. However, they assess child development using the ASQ III, a low-quality measure based on caregiver report. Other studies of parenting interventions that have found sustained impacts two or more years after the end of their programs have thus far either been from HIC settings (Carneiro et al., 2023), featured much lengthier and costly home visiting models of delivery (Yousafzai et al. 2014; 2016; Grantham-McGregor et al. 1991; Gertler et al. 2014), or found sustained impacts after only 15 months and combined across a combination of group and home visit delivery models (Meghir et al. 2023). Like all these other studies, our sustained impacts are smaller than their short-term counterparts, but this is not specific to group delivery: the Lady Health Workers program in Pakistan featured monthly home visits and group sessions over two years and had immediate impacts on child cognition of about 0.6 SD at age 2, but impacts decreased to 0.1-0.2 SD two years later when children were age 4 (Yousafzai et al. 2014; 2016).

Second, our small but marginally positive findings for the booster program extension to help sustain the parental behavioral changes from the original program is overall an encouraging sign and warrants further investigation under less disruptive circumstances. Though the combined *Msingi Bora* program plus boosters results in slightly smaller benefit-cost ratios than the program without boosters, the 6-month interruption of the COVID-19 pandemic affected nearly everything about their implementation, and all our monitoring indicators of engagement and delivery quality experienced an abrupt decline in the wake of the pandemic. Although we lack definitive answers for the underlying causes of these changes, we speculate that factors such as transitioning to online (versus in-person) trainings to our team of trainers, CHVs facing new and increased tasks demanded by local health services in response to the pandemic, and a general shift in priorities among families during this turbulent period may have all contributed to the lower delivery quality of the extended booster sessions. Despite these challenges, we believe some form of light-touch continued program support can play a key role in helping parents to sustain and adapt the behavioral changes over time and for potentially small added costs to an existing program. As such, we recommend further tests of such booster extensions.

Finally, with or without boosters, we also demonstrate significant cost-effectiveness of the program using group-based delivery that was much cheaper than the one-on-one home visiting approaches that have been used previously. Our cost per child of $119 in 2020 US dollars terms is higher than the $76 per child of the First Steps intervention in Rwanda (Justino et al. 2022), but our costs become comparable after adjusting for inflation and converting to purchasing power parity (PPP) terms.[20] Both studies suggest the potential feasibility of scaling group-based parenting interventions in contexts of weak institutions and low levels of human capital. Finally, our benefit-cost ratios of 6.9-7.5 without and with the booster extension are higher than the very few studies reporting similar estimates, including the scaled Reach-up home visiting program in Peru (Araujo et al. 2021), and other health center-based and home visiting programs in Jamaica and other Caribbean countries (Walker et al. 2015); they bookend the 7.1 BCR from the NEP program in Chile when both programs adopt the same 5% discount rate (Carneiro et al. 2023). Overall, our findings add to the small but growing literature documenting the cost-effectiveness and benefit-costs of group-based responsive parenting programs and underscore their higher scalability potential relative to individual home visits, given their higher cost-effectiveness and benefits that remain sustained even years later.

In sum, the results point to encouraging potential paths forward for maximizing the reach and effectiveness of programs to improve children's development in low-resource remote settings such as rural Kenya, where a majority (56%) of children under age 5 are growing up in poverty and subject to numerous risks compromising their development trajectories (Heckman and Mosso 2014; Almond, Currie, and Duque 2018). Future research in this area should include interventions and program delivery systems that further test ways to optimize effectiveness and scalability. For example, there is great potential for the use of digital technology (e.g., mobile phones) to reach families in rural areas who may not have access to community health centers or non-profit organizations. Another key research gap is how to continue to make interventions

[20] The First Steps program from Rwanda was implemented between November 2015 and April 2016, and they do not specify the time period for their quoted costs per child in US dollars. We assume they are in 2016 US dollars.

more sustainable and effective over time. Some approaches that could be tested include to incorporate other family members (e.g., fathers, grandparents) so that children can benefit from multiple nurturing caregivers. Finally, additional research would also be welcome on how best to integrate programs into the existing policy landscape.

References

Abboah-Offei, Mary, Patrick Amboka, Margaret Nampijja, George Evans Owino, Kenneth Okelo, Patricia Kitsao-Wekulo, Ivy Chumo, et al. 2022. "Improving Early Childhood Development in the Context of the Nurturing Care Framework in Kenya: A Policy Review and Qualitative Exploration of Emerging Issues with Policy Makers." *Frontiers in Public Health*, 3499.

Aboud, Frances E, and Sadika Akhter. 2011. "A Cluster-Randomized Evaluation of a Responsive Stimulation and Feeding Intervention in Bangladesh." *Pediatrics* 127 (5): e1191----e1197.

Aboud, Frances E, and Aisha K Yousafzai. 2015. "Global Health and Development in Early Childhood." *Annual Review of Psychology* 66: 433–57.

Almeida, David M, Elaine Wethington, and Ronald C Kessler. 2002. "The Daily Inventory of Stressful Events: An Interview-Based Approach for Measuring Daily Stressors." *Assessment* 9 (1): 41–55.

Araujo, M Caridad, Marta Dormal, Sally Grantham-McGregor, Fabiola Lazarte, Marta Rubio-Codina, and Norbert Schady. 2021. "Home Visiting at Scale and Child Development." *Journal of Public Economics Plus* 2: 100003.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020. "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia." *American Economic Review* 110 (1): 48–85.

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020. "Human Capital Development and Parental Investment in India." *The Review of Economic Studies* 87 (6): 2511–41.

Attanasio, Orazio, Costas Meghir, Emily Nix, and Francesca Salvati. 2017. "Human Capital Growth and Poverty: Evidence from Ethiopia and Peru." *Review of Economic Dynamics* 25: 234–59. https://doi.org/10.1016/j.red.2017.02.002.

Bandura, Albert. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Prentice-Hall, Inc.

———. 1995. *Self-Efficacy in Changing Societies*. Cambridge University Press.

Beguy, Donatien, Philippe Bocquier, and Eliya Msiyaphazi Zulu. 2010. "Circular Migration Patterns and Determinants in Nairobi Slum Settlements." *Demographic Research* 23 (20): 549–86. https://www.demographic-research.org/volumes/vol23/20/.

Black, Maureen M, Susan P Walker, Lia C H Fernald, Christopher T Andersen, Ann M DiGirolamo, Chunling Lu, Dana C McCoy, et al. 2017. "Early Childhood Development Coming of Age: Science through the Life Course." *The Lancet* 389 (10064): 77–90. https://doi.org/https://doi.org/10.1016/S0140-6736(16)31389-7.

Boivin, Michel, Daniel Pérusse, Ginette Dionne, Valérie Saysset, Mark Zoccolillo, George M Tarabulsy, Nathalie Tremblay, and Richard E Tremblay. 2005. "The Genetic-Environmental Etiology of Parents' Perceptions and Self-Assessed Behaviours toward Their 5-Month-Old Infants in a Large Twin and Singleton Sample." *Journal of Child Psychology and Psychiatry* 46 (6): 612–30.

Bretherton, Inge. 1992. "The Origins of Attachment Theory: John Bowlby and Mary Ainsworth." *Developmental Psychology* 28 (5): 759.

Britto, Pia R, Stephen J Lye, Kerrie Proulx, Aisha K Yousafzai, Stephen G Matthews, Tyler Vaivada, Rafael Perez-Escamilla, et al. 2017. "Nurturing Care: Promoting Early Childhood Development." *The Lancet* 389 (10064): 91–102.

Caldwell, Bettye M, and Robert H Bradley. 2003. "Home Inventory Administration Manual." University of Arkansas for Medical Sciences.

Campos, Francisco, Michael Frese, Markus Goldstein, Leonardo Iacovone, Hillary C. Johnson, David McKenzie, and Mona Mensmann. 2017. "Teaching Personal Initiative Beats Traditional Training in Boosting Small Business in West Africa." *Science* 357 (6357): 1287–90. https://doi.org/10.1126/science.aan5329.

Carneiro, P., C. Meghir, and M Parey. 2013. "Maternal Education, Home Environments, and the Development of Children and Adolescents." *Journal of the European Economic Association* 11 (s1): 123–60.

Carneiro, Pedro, Emanuela Galasso, Italo Lopez Garcia., Miguel Cordero, and Paula Bedregal. 2023. "Impacts of a Large-Scale Parenting Program: Experimental Evidence from Chile." *Journal of Political Economy (Forthcoming)*.

Chernozhukov, Victor, Iván Fernández-Val, and Ye Luo. 2018. "The Sorted Effects Method: Discovering Heterogeneous Effects beyond Their Averages." *Econometrica* 86 (6): 1911–38.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97 (2): 31–47.

Cunha, Flavio, James J Heckman, and Susanne M Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.

DHS. 2009. "Kenya Demographic and Health Survey 2008-2009 Final Report."

Doyle, Orla. 2020. "The First 2,000 Days and Child Skills." *Journal of Political Economy* 128 (6): 2067–2122.

Duggan, Cian, Alan D Irvine, Jonathan O'B Hourihane, Mairead E Kiely, and Deirdre M Murray. 2023. "ASQ-3 and BSID-III's Concurrent Validity and Predictive Ability of Cognitive Outcome at 5 Years." *Pediatric Research*, 1–7.

Engle, Patrice L, Lia C H Fernald, Harold Alderman, Jere Behrman, Chloe O'Gara, Aisha Yousafzai, Meena Cabral de Mello, et al. 2011. "Strategies for Reducing Inequalities and Improving Developmental Outcomes for Young Children in Low-Income and Middle-Income Countries." *The Lancet* 378 (9799): 1339–53.

Fernald, Lia C, Lynnette M Neufeld, Lauren R Barton, Lourdes Schnaas, Juan Rivera, and Paul J Gertler. 2006. "Parallel Deficits in Linear Growth and Mental Development in Low-Income Mexican Infants in the Second Year of Life." *Public Health Nutrition* 9 (2): 178–86. https://doi.org/DOI: 10.1079/PHN2005759.

Fiorini, Mario, and Michael P Keane. 2014. "How the Allocation of Children's Time Affects Cognitive and Noncognitive Development." *Journal of Labor Economics* 32 (4): 787–836.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor. 2014. "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica." *Science* 344 (6187): 998–1001.

Goodman, Robert. 1997. "The Strengths and Difficulties Questionnaire: A Research Note." *Journal of Child Psychology and Psychiatry* 38 (5): 581–86.

Grantham-McGregor, Sally, Akanksha Adya, Orazio Attanasio, Britta Augsburg, Jere Behrman, Bet Caeyers, Monimalika Day, et al. 2020. "Group Sessions or Home Visits for Early Childhood Development in India: A Cluster RCT." *Pediatrics* 146 (6).

Grantham-McGregor, Sally M, Christine A Powell, Susan P Walker, and John H Himes. 1991. "Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study." *The Lancet* 338 (8758): 1–5.

Hamadani, Jena D, Fahmida Tofail, Afroza Hilaly, Syed N Huda, Patrice Engle, and Sally 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.

Hanlon, C, G Medhin, B Worku, M Tomlinson, A Alem, M Dewey, and M Prince. 2016. "Adapting the Bayley Scales of Infant and Toddler Development in Ethiopia: Evaluation of Reliability and Validity." *Child: Care, Health and Development* 42 (5): 699–708.

Heckman, James J. 2023. "Jamaica Reach Up: Its Contribution to Policy and Science." *Pediatrics* 151 (Supplement 2).

Heckman, James J, Margaret L Holland, Kevin K Makino, Rodrigo Pinto, and Maria Rosales-Rueda. 2017. "An Analysis of the Memphis Nurse-Family Partnership Program." *NBER Working Paper*, no. 23610.

Jensen, Sarah K G, Matias Placencio-Castro, Shauna M Murray, Robert T Brennan, Simo Goshev, Jordan Farrar, Aisha Yousafzai, et al. 2021. "Effect of a Home-Visiting Parenting Program to Promote Early Childhood Development and Prevent Violence: A Cluster-Randomized Trial in Rwanda." *BMJ Global Health* 6 (1): e003508.

Jeong, Joshua, Emily E Franchett, Clariana V de Oliveira, Karima Rehmani, and Aisha K Yousafzai. 2021. "Parenting Interventions to Promote Early Child Development in the First Three Years of Life: A Global Systematic Review and Meta-Analysis." *PLoS Medicine* 18 (5): e1003602.

Jeong, Joshua, Helen O Pitchik, and Günther 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): e004067.

Jervis, Pamela, Jacqueline Coore-Hall, Helen O Pitchik, Charles D Arnold, Sally Grantham-McGregor, Marta Rubio-Codina, Helen Baker-Henningham, Lia C H Fernald, Jena Hamadani, and Joanne A Smith. 2023. "The Reach Up Parenting Program, Child Development, and Maternal Depression: A Meta-Analysis." *Pediatrics* 151 (Supplement 2).

Justino, Patricia, Marinella Leone, Pierfrancesco Rolla, Monique Abimpaye, Caroline Dusabe, Marie D Uwamahoro, and Richard Germond. 2022. "Improving Parenting Practices for Early Child Development: Experimental Evidence from Rwanda." *Journal of the European Economic Association*, jvac070.

Kalil, Ariel, and Rebecca Ryan. 2020. "Parenting Practices and Socioeconomic Gaps in Childhood Outcomes." *The Future of Children* 30 (1): 29–54.

Kenya Ministry of Health. 2020. "Kenya Community Health Policy 2020-2030." Nairobi, Kenya.

Knauer, Heather A, Patricia Kariger, Pamela Jakiela, Owen Ozier, and Lia C H Fernald. 2019. "Multilingual Assessment of Early Child Development: Analyses from Repeated Observations of Children in Kenya." *Developmental Science* 22 (5): e12875.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76 (3): 1071–1102.

Lopez Garcia, Italo, Uzaib Saya, and Jill E Luoto. 2021. "Cost-Effectiveness and Economic Returns of Group-Based Parenting Interventions to Promote Early Childhood Development: Results from a Randomized Controlled Trial in Rural Kenya." *PLoS Medicine*, September. https://doi.org/https://doi.org/10.1371/ journal.pmed.1003746.

Lu, Chunling, Maureen M Black, and Linda M Richter. 2016. "Risk of Poor Development in Young Children in Low-Income and Middle-Income Countries: An Estimation and Analysis at the Global, Regional, and Country Level." *The Lancet Global Health* 4 (12): e916–e922.

Lubben, James, Eva Blozik, Gerhard Gillmann, Steve Iliffe, Wolfgang von Renteln Kruse, John C Beck, and Andreas E Stuck. 2006. "Performance of an Abbreviated Version of the Lubben Social Network Scale among Three European Community-Dwelling Older Adult Populations." *The Gerontologist* 46 (4): 503–13.

Luoto, Jill E, Italo Lopez Garcia, Frances E Aboud, Daisy R Singla, Lia C H Fernald, Helen O Pitchik, Uzaib Y Saya, Ronald Otieno, and Edith Alu. 2020. "Group-Based Parenting Interventions to Promote Child Development in Rural Kenya: A Multi-Arm, Cluster-Randomised Community Effectiveness Trial." *The Lancet Global Health* 9 (3): e309–19.

Luoto, Jill E, Italo Lopez Garcia, Frances E Aboud, Daisy R Singla, Rebecca Zhu, Ronald Otieno, and Edith Alu. 2021. "An Implementation Evaluation of A Group-Based Parenting Intervention to Promote Early Childhood Development in Rural Kenya." *Frontiers in Public Health* 9 (May): 653106. https://doi.org/10.3389/fpubh.2021.653106.

Meghir, Costas, Orazio Attanasio, Pamela Jervis, Monimalika Day, Prerna Makkar, Jere Behrman, Rashim Pal, Angus Phimister, Nisha Vernekar, and Sally Grantham-McGregor. 2023. "Early Stimulation and Enhanced Preschool: A Randomized Trial." *Pediatrics* 151 (Supplement 2).

Pendergast, Laura L, Barbara A Schaefer, Laura E Murray-Kolb, Erling Svensen, Rita Shrestha, Muneera A Rasheed, Rebecca J Scharf, et al. 2018. "Assessing Development across Cultures: Invariance of the Bayley-III Scales Across Seven International MAL-ED Sites." *School Psychology Quarterly* 33 (4): 604.

Rademeyer, Vanessa, and Lorna Jacklin. 2013. "A Study to Evaluate the Performance of Black South African Urban Infants on the Bayley Scales of Infant Development III." *South African Journal of Child Health* 7 (2): 54–59.

Radloff, Lenore Sawyer. 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1 (3): 385–401.

Radner, James M, Marvin J S Ferrer, Dominique McMahon, Anuraj H Shankar, and Karlee L Silver. 2018. "Practical Considerations for Transitioning Early Childhood Interventions to Scale: Lessons from the Saving Brains Portfolio." *Annals of the New York Academy of Sciences* 1419 (1): 230–48.

Raiford, Susan Engi, and Diane L Coalson. 2014. *Essentials of WPPSI-IV Assessment*. John Wiley & Sons.

Rijen, E H M Van, N Gasanova, A M Boonstra, and Jorg Huijding. 2014. "Psychometric Qualities of the Short Form of the Self-Efficacy for Parenting Tasks Index-Toddler Scale." *Child Psychiatry & Human Development* 45 (4): 443–55.

Romano, Joseph P, and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.

Rubio-Codina, Marta, M Caridad Araujo, Orazio Attanasio, Pablo Muñoz, and Sally Grantham-McGregor. 2016. “Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies.” *PloS One* 11 (8): e0160962.

Singla, Daisy R, Elias Kumbakumba, and Frances E Aboud. 2015. “Effects of a Parenting Intervention to Address Maternal Psychological Wellbeing and Child Development and Growth in Rural Uganda: A Community-Based, Cluster-Randomised Trial.” *The Lancet Global Health* 3 (8): e458----e469.

Tomlinson, M, X Hunt, and MJ Rotheram-Borus. 2018. “Diffusing and Scaling Evidence-Based Interventions: Eight Lessons for Early Child Development from the Implementation of Perinatal Home Visiting in South Africa.” *Annals of the New York Academy of Sciences* 1419 (1): 218–29. https://doi.org/10.1111/nyas.13650.

Tucker-Drob, Elliot M. 2009. “Differentiation of Cognitive Abilities across the Life Span.” *Developmental Psychology* 45 (4): 1097.

Veldhuizen, Scott, Jean Clinton, Christine Rodriguez, Terrance J Wade, and John Cairney. 2015. “Concurrent Validity of the Ages and Stages Questionnaires and Bayley Developmental Scales in a General Population Sample.” *Academic Pediatrics* 15 (2): 231–37.

Walker, Susan P, Christine Powell, Susan M Chang, Helen Baker-Henningham, Sally Grantham-McGregor, Marcos Vera-Hernandez, and Florencia López-Boo. 2015. “Delivering Parenting Interventions through Health Services in the Caribbean: Impact, Acceptability and Costs.”

Willoughby, Michael T, Benjamin Piper, Dunston Kwayumba, and Megan McCune. 2019. “Measuring Executive Function Skills in Young Children in Kenya.” *Child Neuropsychology* 25 (4): 425–44.

Wolke, Dieter, David Skuse, and Berenice Mathisen. 1990. “Behavioral Style in Failure-to-Thrive Infants: A Preliminary Communication.” *Journal of Pediatric Psychology* 15 (2): 237–54.

World Health Organization. 2018. “Nurturing Care for Early Childhood Development: A Framework for Helping Children Survive and Thrive to Transform Health and Human Potential.”

World Health Organization, and UNICEF. 2003. *Global Strategy for Infant and Young Child Feeding*.

Yousafzai, Aisha K, and Frances Aboud. 2014. “Review of Implementation Processes for Integrated Nutrition and Psychosocial Stimulation Interventions.” *Annals of the New York Academy of Sciences* 1308 (1): 33–45.

Yousafzai, Aisha K., Jelena Obradović, Muneera A. Rasheed, Arjumand Rizvi, Ximena A. Portilla, Nicole Tirado-Strayer, Saima Siyal, and Uzma Memon. 2016. “Effects of Responsive Stimulation and Nutrition Interventions on Children’s Development and Growth at Age 4 Years in a Disadvantaged Population in Pakistan: A Longitudinal Follow-up of a Cluster-Randomised Factorial Effectiveness Trial.” *The Lancet Global Health* 4 (8): e548–58. https://doi.org/10.1016/S2214-109X(16)30100-0.

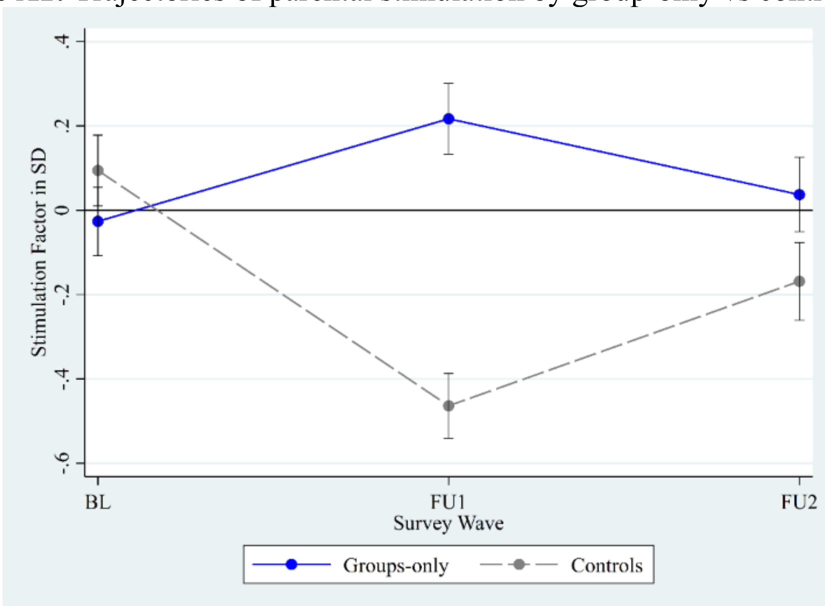
Yousafzai, Aisha K, Muneera A Rasheed, Arjumand Rizvi, Robert Armstrong, and Zulfiqar A Bhutta. 2014. “Effect of Integrated Responsive Stimulation and Nutrition Interventions in the Lady Health Worker Programme in Pakistan on Child Development, Growth, and Health Outcomes: A Cluster-Randomised Factorial Effectiveness Trial.” *The Lancet* 384 (9950): 1282–93.

Yue, Ai, Qi Jiang, Biaoyue Wang, Cody Abbey, Alexis Medina, Yaojiang Shi, and Scott Rozelle. 2019. "Concurrent Validity of the Ages and Stages Questionnaire and the Bayley Scales of Infant Development III in China." *PLoS One* 14 (9): e0221675.

Appendix

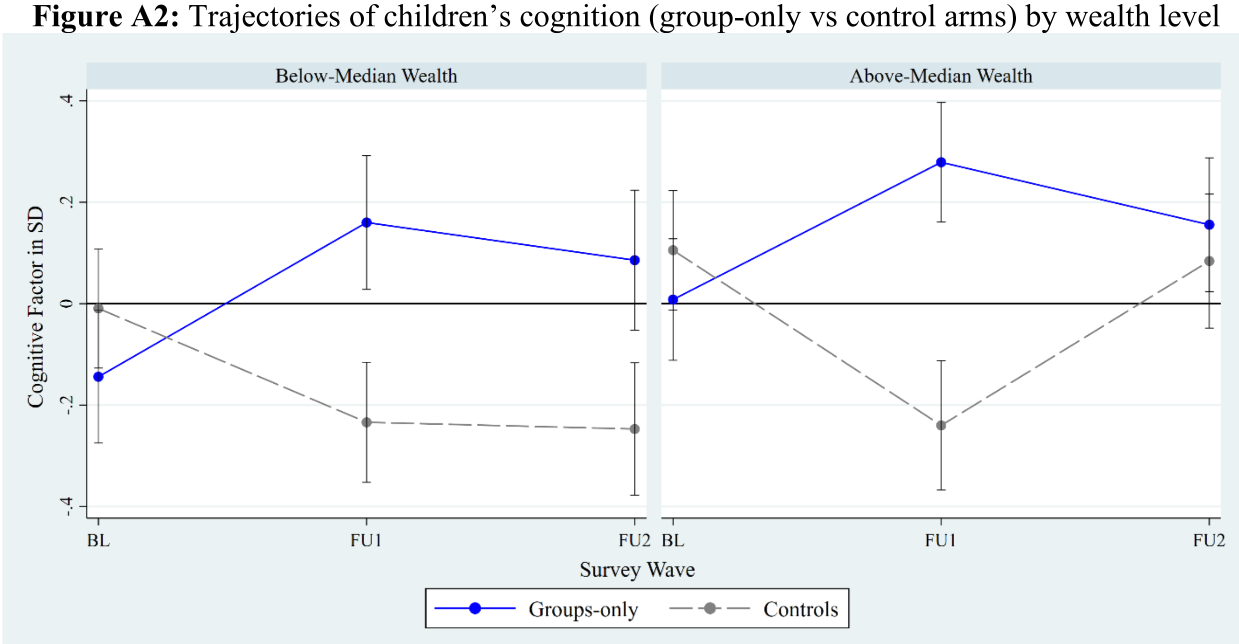
Appendix A: Figures and Tables

Figure A1: Trajectories of parental stimulation by group-only vs control arms



Notes: Confidence intervals at the 90% level

Figure A2: Trajectories of children's cognition (group-only vs control arms) by wealth level



Notes: Confidence intervals at the 90% level

Table A1: Baseline characteristics and sample balance between control group, mixed-delivery and group-only (reprinted from Luoto et al. (2020))

	(1) Control group	(2) Group-only	(3) Mixed-delivery	(4) Pairwise group differences (p-value) 2 vs. 1	(5) 3 vs. 1	(6) 3 vs.2	(7) Joint F-test (p-value)
Panel A. Continuous variables							
Mother's age	29.19	28.29	28.17	0.284	0.217	0.884	0.405
	(9.00)	(9.07)	(8.26)				
Mother's education (years)	8.97	8.93	8.80	0.838	0.435	0.576	0.723
	(2.51)	(2.65)	(2.75)				
Household size	5.58	5.51	5.86	0.675	0.129	0.066	0.156
	(1.89)	(2.13)	(2.73)				
Wealth index	-0.06	0.05	0.01	0.156	0.392	0.634	0.312
	(0.92)	(1.01)	(1.04)				
Child age (months)	14.20	13.86	14.43	0.465	0.602	0.175	0.389
	(4.66)	(5.02)	(4.67)				
Child's Cognitive Factor	0.00	-0.12	-0.03	0.310	0.814	0.442	0.563
	(0.99)	(1.04)	(1.04)				
HOME Stimulation Factor	0.00	-0.11	-0.16	0.324	0.192	0.751	0.345
	(0.98)	(1.04)	(0.97)				
Panel B. Categorical variables							
Father in household (%)	61.01	60.59	64.00	0.912	0.476	0.386	0.657
Female child (%)	51.19	47.33	50.75	0.289	0.909	0.341	0.449
Birth order				0.208	0.225	0.481	
First (%)	21.75	27.01	27.50				0.155
Second or third (%)	21.75	22.73	19.25				0.522
Fourth or more (%)	56.50	50.27	53.25				0.279

Joint F-test (p-value)				0.174	0.034	0.571
Observations	377	374	400	750	777	773
Number of villages (clusters)	20	20	20	400	400	400

Notes: Data are means (SD) or frequency (%), unless otherwise stated. Standard deviations are clustered at the village level. Bayley scores are scaled scores age-standardized (range 0 to 19) using the publisher's manual. Center for Epidemiologic Studies Depression Scale (CES-D) scores at baseline are on a 0-40 scale due to an error in scoring at baseline and are not comparable with scores at endline or to outside samples. *Family Care Indicator (FCI) scores for fathers at baseline have N=162 for comparison arm, N=160 for group-only arm, N=190 for mixed-delivery arm, N=193 for father villages, and N=157 for mother-only villages. Child length-for-age at baseline used Seca mobile measuring mats (model 210). Enumerators measured the child three times and calculated the average. All measures were converted to length-for-age Z-scores following WHO recommendations and stunting was defined as <2 SD below the mean.

Table A2: Sample Attrition and Intervention Status

	(1) Basic Model	(2) Adjusted Model	(3) Full Interactions
Treated, no boosters	-0.002	0.008	0.238
	(0.031)	(0.033)	(0.191)
Treated + Boosters	-0.012	-0.010	-0.289
	(0.025)	(0.027)	(0.149)
South Rachuonyo	0.077*	0.057	0.057
	(0.031)	(0.033)	(0.033)
East Rachuonyo	-0.014	-0.032	-0.034
	(0.026)	(0.030)	(0.031)
Mother's age		-0.006***	-0.008***
		(0.001)	(0.002)
Mother's years of schooling		0.004	-0.012
		(0.005)	(0.010)
Wealth index at baseline		-0.022	-0.013
		(0.015)	(0.029)
Child age at baseline (months)		-0.003	-0.006
		(0.002)	(0.004)
Female child (%)		-0.004	0.020
		(0.022)	(0.040)
Birth order 2		-0.034	0.004
		(0.038)	(0.057)
Birth order 3+		-0.100**	-0.093
		(0.031)	(0.051)
Baseline child cognitive ability (age-std.)		-0.010	-0.030
		(0.011)	(0.019)
Treated, no boosters X Mother's age			0.005
			(0.003)
Treated + Boosters X Mother's age			0.002
			(0.003)
Treated, no boosters X Mother's years of schooling			0.032*
			(0.012)
Treated + Boosters X Mother's years of schooling			0.015
			(0.013)
Treated, no boosters X Wealth index at baseline			-0.025
			(0.039)
Treated + Boosters X Wealth index at baseline			-0.004
			(0.037)
Treated, no boosters X Child age at baseline (months)			0.008
			(0.006)
Treated + Boosters X Child age at baseline (months)			0.000
			(0.006)
Female X Treated, no boosters=1			-0.028
			(0.049)
Female X Treated + Boosters=1			-0.060
			(0.057)
Birth order 2=1 X Treated, no boosters=1			-0.055

	(1)	(2)	(3)
			(0.074)
Birth order 2=1 X Treated + Boosters=1			-0.050
			(0.097)
Birth order 3+=1 X Treated, no boosters=1			0.015
			(0.071)
Birth order 3+=1 X Treated + Boosters=1			-0.029
			(0.075)
Treated, no boosters X Baseline child cognitive ability (age-std.)			0.025 (0.025)
Treated + Boosters X Baseline child cognitive ability (age-std.)			0.031 (0.026)
Observations	1152	1152	1152
F-test P-value joint significance treatment variables	0.884	0.852	0.343

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses clustered at the village. Results from linear probability models predicting attrition at two-year follow-up survey. Treated w/o boosters takes on a value of 1 if a mother-child dyad received the group-only or mixed-delivery intervention during first phase of *Msingi Bora* and no subsequent boosters. Treated + boosters takes on a value of 1 if a dyad received group-only or mixed-delivery interventions followed by the booster extension. Balanced attrition at Endline across the two delivery models was demonstrated in Luoto et al. (2020) and is not repeated here.

Table A3: ITT Impacts on Child Cognitive & Socioemotional individual indices at Follow-Up2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	BD total	Total Receptive Language	Total Expressive Language	Executive Functioning index	Total Wolke Score	SDQ: Prosocial Behavior	SDQ: Peer problems	SDQ: Hyperactivity problems	SDQ: Emotional Symptoms	SDQ: Conduct Problems
1. Groups No Boosters	0.159*	0.034	0.110	0.114	0.161**	0.050	0.121	-0.045	0.163***	0.133
	(0.095)	(0.101)	(0.098)	(0.084)	(0.074)	(0.082)	(0.123)	(0.100)	(0.058)	(0.106)
2. Groups + Boosters	0.241**	0.130	-0.059	0.126	0.203***+	0.050	0.258***	0.033	0.003	0.231***
	(0.106)	(0.082)	(0.089)	(0.091)	(0.075)	(0.092)	(0.103)	(0.111)	(0.099)	(0.098)
3. Mixed No Boosters	0.061	-0.008	-0.049	0.003	0.237***+	0.138	0.102	-0.141**	-0.088	-0.052
	(0.102)	(0.082)	(0.126)	(0.137)	(0.082)	(0.106)	(0.083)	(0.064)	(0.102)	(0.097)
4. Mixed + Boosters	0.068	0.065	-0.004	0.026	0.035	0.150*	0.193**	0.190**	0.090	0.136
	(0.096)	(0.113)	(0.070)	(0.106)	(0.074)	(0.086)	(0.089)	(0.096)	(0.119)	(0.098)
Observations	934	934	934	917	934	941	941	941	941	941
p-value Boosters>No Boosters										
Overall	0.269	0.140	0.787	0.418	0.938	0.461	0.087	0.006	0.459	0.019
Group	0.220	0.185	0.944	0.442	0.268	0.500	0.161	0.277	0.934	0.161
Mixed delivery	0.471	0.268	0.349	0.440	0.994	0.450	0.167	0.000	0.115	0.025

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Standard errors in parentheses are clustered at the village level. Each column is a separate regression for each outcome. Parental outcomes in columns 2-5. + Signifies estimate is significant at 10% level using Romano-Wolf estimator that corrects for multiple hypothesis testing. ++ Signifies estimate significant at 5% level using Romano-Wolf estimator. † Signifies estimate significant at 1% level using Romano-Wolf estimator. Romano-Wolf p-values were computed for two groups/families of estimations: 1. columns (1) to (5) and 2. columns (6) to (10). Data are in SD. Effect sizes and p-values obtained from intention-to-treat (ITT) estimates for each outcome measured at endline using internal age-standardization to the comparison group. All results include adjustments prespecified in our study protocol: child's age, household wealth, maternal education, child sex, birth order, the outcome at baseline (if measured) and sub-county fixed effects (the strata).

Table A4: Lee Bounds for Impacts on Child Cognitive and Socioemotional Indices at Follow-Up 2

	(1)	(2)	(3)	(4)
	Groups, No Boosters	Groups + Boosters	Mixed, No Boosters	Mixed + Boosters

1. Cognitive Index				
Lower Bound	0.137	0.230**	0.032	0.098
	[-0.052,0.327]	[0.027,0.433]	[-0.195,0.259]	[-0.133,0.328]
Upper Bound	0.244*	0.244**	0.058	0.117
	[-0.012,0.499]	[0.038,0.450]	[-0.170,0.285]	[-0.115,0.349]
Observations	564	564	568	586
2. Socioemotional Index				
Lower Bound	0.170*	0.223**	0.036	0.222**
	[-0.010,0.350]	[0.024,0.421]	[-0.144,0.215]	[0.020,0.425]
Upper Bound	0.254***	0.258**	0.036	0.250**
	[0.066,0.442]	[0.049,0.466]	[-0.155,0.226]	[0.050,0.450]
Observations	564	564	568	586

Notes: * p<0.10, ** p< 0.05, *** p<0.01. Child cognition in upper panel. Socioemotional in lower panel. 95% confidence intervals (displayed in square brackets) computed using cluster-bootstrap standard errors at the village level.

Table A5: Lee Bounds for Impacts on Parental Behavioral Indices at Follow-Up 2

	(1) Groups, No Boosters	(2) Groups + Boosters	(3) Mixed, No Boosters	(4) Mixed + Boosters
1. Stimulation Index				
Lower Bound	0.120	0.229	0.190	0.387***
	[-0.081,0.321]	[-0.052,0.510]	[-0.043,0.424]	[0.146,0.628]
Upper Bound	0.252*	0.284*	0.245*	0.397***
	[-0.006,0.510]	[-0.012,0.581]	[-0.030,0.519]	[0.147,0.647]
Observations	564	564	568	586
2. Knowledge Index				
Lower Bound	0.294***	0.357***	0.157	0.241*
	[0.094,0.493]	[0.124,0.591]	[-0.104,0.417]	[-0.017,0.500]
Upper Bound	0.381***	0.364***	0.172	0.258*
	[0.194,0.568]	[0.123,0.604]	[-0.109,0.452]	[-0.001,0.517]
Observations	564	564	568	586

Notes: * p<0.10, ** p< 0.05, *** p<0.01. HOME Stimulation in upper panel. Knowledge in lower panel. 95% confidence intervals (displayed in square brackets) computed using cluster-bootstrap standard errors at the village level.

Table A6: Individual and Grouped estimation of the Mediation Effects on Children's Cognitive Development at Follow-up 2

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Children's cognition at FU2	Benchmark	Cognition at FU1	Stimulation at FU2	Knowledge at FU2	Nutrition at FU2	+ Knowledge FU2
Groups + Boosters	0.228**	0.100	0.170**	0.177*	0.233***	0.017
	(0.088)	(0.082)	(0.084)	(0.091)	(0.087)	(0.081)
Groups + Non-boosters	0.200**	0.112	0.155*	0.154	0.200*	0.044
	(0.099)	(0.100)	(0.090)	(0.102)	(0.101)	(0.096)
Children's Cognition at FU1		0.192***				0.174***
		(0.030)				(0.031)

	(1)	(2)	(3)	(4)	(5)	(6)
Stimulation at FU2			0.262***			0.235***
			(0.042)			(0.039)
Knowledge at FU2				0.139***		0.111***
				(0.032)		(0.034)
Nutrition at FU2					-0.082	
					(0.070)	
Observations	945	917	943	945	942	915
Indirect Effect Groups+Boosters		**56.1%**	**25.4%**	**22.4%**	-2.2%	**92.5%**
Indirect Effect Groups+No Boosters		**44.0%**	**22.5%**	**23.0%**	0.0%	**78.0%**

Table A7: Individual and Grouped estimation of the Mediation Effects on Children's Socioemotional Development at Follow-up 2

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.: Children's socioemotional at FU2	Benchmark	Cognition at FU1	Stimulation at FU2	Knowledge at FU2	Nutrition at FU2	+ Knowledge FU2
Groups + Boosters	0.233***	0.170**	0.123*	0.205**	0.222**	0.063
	(0.084)	(0.084)	(0.063)	(0.085)	(0.084)	(0.065)
Groups + Non-boosters	0.204**	0.163*	0.115	0.179**	0.197**	0.080
	(0.082)	(0.086)	(0.094)	(0.086)	(0.080)	(0.097)
Children's Cognition at FU1		0.091***				0.058**
		(0.029)				(0.028)
Stimulation at FU2			0.450***			0.444***
			(0.029)			(0.031)
Knowledge at FU2				0.075**		0.041
				(0.034)		(0.034)
Nutrition at FU2					0.104	
					(0.077)	
Observations	945	917	943	945	942	915
Indirect Effect Groups+Boosters		**27.0%**	**47.2%**	**12.0%**	4.7%	**73.0%**
Indirect Effect Groups+Non Boosters		**20.1%**	**43.6%**	**12.3%**	3.4%	**60.8%**

Appendix B: Benefit-Cost Analyses Details

We translate our two-year impacts on children's cognitive and socioemotional development at ages 3.5-5 to impacts on adult schooling and earnings outcomes in two steps. First, we multiply our two-year ITT impacts from Table 4 in the manuscript with estimates of the wage and educational returns to cognitive impacts obtained from rounds 1-4 of the representative Kenya Life Panel Survey (KLPS) which contains longitudinal data on cognitive tests, schooling, earnings, and household sociodemographics. Using multivariate linear regressions of log wages on an age-standardized measure of cognition, we predict that a one SD increase in children's cognition is associated with a 39.7% increase in annual wages after controlling for parental education, survey wave and month of interview, a female indicator variable, and baseline school grade fixed effects. Using a similar estimation, we predict a one SD increase in children's cognition is associated with an additional 1.79 years of schooling. However, the KLPS does not include wage or schooling returns to socioemotional skills. We instead use an estimate of a 6.4% wage return to socioemotional skills based on a recent study from Kenya (Otchia and Yamada 2019); for schooling returns, we assume the same returns for socioemotional as for cognitive improvements and again use the KLPS. For simplicity, we assume that the total program impact on adult schooling or wages is just the sum of impacts stemming from cognitive and socioemotional skill improvements.

Next, to calculate the gains in cumulative lifetime wages using these estimated returns to wages and schooling, we obtain wage profiles by age of children as adults using the 2015-2016 nationally representative Kenya Integrated Household Budget Survey (KIHBS). We restrict the sample to at least halftime workers (working at least 20 hours per week in any paid occupation) and calculate the life stream of average earnings by age expressed in 2020 prices. We convert the sum of discounted earnings to present values adjusting for expected survival probabilities using age-life tables from Kenya and a discount factor of 5% in order to be consistent with our results estimated based on our short-term results as published previously (Lopez Garcia et al., 2021).

Finally, to calculate the long-term additional schooling costs we multiply our predicted intervention impacts on schooling with the public cost of an additional grade per child per year, which we assume to be 15% of the basic wage rate for low-skilled workers based on KIHBS wage data, following Nandi et al. (2017) as well as our short-term analysis (Lopez Garcia et al., 2021).