NBER WORKING PAPER SERIES

EARLY STIMULATION AND NUTRITION:
THE IMPACTS OF A SCALABLE INTERVENTION

Orazio Attanasio
Helen Baker-Henningham
Raquel Bernal
Costas Meghir
Diana Pineda
Marta Rubio-Codina

Working Paper 25059
http://www.nber.org/papers/w25059

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2018, Revised April 2021

This study was funded by Grand Challenges Canada Grant 0462-03-10 and Fundación Éxito. Orazio Attanasio was partly funded by ERC Advanced Grant AdG – 695300. Costas Meghir was partly funded by the NIH grant R01HD7210, the Cowles Foundation and ISPS at Yale. This trial is registered at the ISRCTN Registry, trial number ISRCTN93757590. The Universidad de los Andes ethics committee (No 287/2014) and the University College London research ethics committee (No 2168/011) approved this study. We would like to thank the Instituto Colombiano de Bienestar Familiar (ICBF), ICBF program supervisors in Cundinamarca, Santander and Boyacá and program coordinators at Fundación Éxito for their cooperation in facilitating the intervention; the FAMI program providers, children and families who willingly participated in this study; all the study staff including our nine tutors and María Lucía Gómez, our field manager; all the research staff including Santiago Lacouture, Alejandro Sánchez, Sara Ramírez and Diana Pérez; the data collection team from IQuartil; and the experts from Grand Challenges Canada. We received useful comments from the editor and 3 anonymous referees. The views here presented are those of the authors and do not represent those of the institutions they represent, including those of the Inter-American Development Bank, its board of directors, or the countries they represent. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Early Stimulation and Nutrition: The Impacts of a Scalable Intervention
Orazio Attanasio, Helen Baker-Henningham, Raquel Bernal, Costas Meghir, Diana Pineda, and Marta Rubio-Codina
NBER Working Paper No. 25059
September 2018, Revised April 2021
JEL No. H43,I10,I20,J13

ABSTRACT

Early Childhood Development is becoming the focus of policy worldwide. However, the evidence on the effectiveness of scalable models is scant, particularly when it comes to infants in developing countries. In this paper we describe and evaluate with a cluster-RCT an intervention designed to improve the quality of child stimulation within the context of an existing parenting program in Colombia, known as FAMI. The intervention improved children's development by 0.16 of a standard deviation (SD) and children's nutritional status, as reflected in a reduction of 5.8 percentage points of children whose height-for-age is below - 1 SD.

Orazio Attanasio
Department of Economics
Yale University
37 Hillhouse Avenue
New Haven, CT 06511
and Institute for Fiscal Studies, FAIR,
BREAD and CEPR
and also NBER
orazio.attanasio@yale.edu

Helen Baker-Henningham
Bangor University
School of Psychology
Brigantia Building
Bangor, Gwynedd
Wales, LL57 2AS
United Kingdom
and University of the West Indies
h.henningham@bangor.ac.uk

Raquel Bernal
Universidad de los Andes
and CEDE
rbernal@uniandes.edu.co

Costas Meghir
Department of Economics
Yale University
37 Hillhouse Avenue
New Haven, CT 06511
and IZA
and also NBER
c.meghir@yale.edu

Diana Pineda
Fundación Éxito
dpinedar@Grupo-Exito.com

Marta Rubio-Codina
Inter-American Development Bank
Social Protection and Health Division
Washington, DC 20577
martarubio@iadb.org

1. Introduction

Human capital, important as it is for life outcomes (Becker, 1964) and economic development, is undermined by poverty from the very beginning of life. This, in turn, leads to a vicious cycle: the underachievement of individuals from deprived backgrounds contributes to the intergenerational persistence of poverty. It is now widely understood that the early years of brain development, and indeed the first 1000 days, can be particularly important for adult outcomes, with the experiences during early childhood having a long-lasting impact.[1]

Over the last couple of decades, our understanding of the process of child development and the evidence on the types of interventions that might improve outcomes have advanced significantly (Black et al., 2017; Britto et al., 2017). In particular, the potential of parenting support programs to improve child development, especially in vulnerable contexts, has been amply demonstrated (Neville, Pakulak, and Stevens, 2015; Britto et al., 2017).

Given the established knowledge, early years interventions should aim at improving the ability of parents to provide responsive and emotionally supportive environments and ensure developmentally stimulating opportunities for their children (Bradley, and Putnick, 2012; Singla, Kumbakumba, and Aboud, 2015; Black et al., 2017), while at the same time be implementable at realistic cost levels and given the available implementation infrastructure, including personnel. If well-designed and adequately targeted to the appropriate age and population subgroups, these programs may be crucial in breaking the intergenerational transmission of poverty.

Indeed, governments around the world have recognized the importance of the early years and have started to introduce services to support children from deprived backgrounds. Head Start in the US and Sure Start in the UK are prime examples in developed economies, while the Cuna Más in Peru and the Family, Women and Childhood program (FAMI for its acronym in Spanish) in Colombia, which is our focus in this work, are similar examples in low- or middle-income countries (LMICs). Indeed, an increasing number of countries now have national early childhood policies (Devercelli, Sayre, and Denboba, 2016).

Although at-scale early years programs are becoming widespread, evidence on their long-term effectiveness—that is, their ability to improve early childhood development (ECD) outcomes in a manner that translates into improved functioning and well-being later in life— is limited. Long run impacts will likely vary depending on the detail of *what* they actually offer and *how* they actually offer it. Understanding their effectiveness in the context of LMICs is even more important than in high-income countries as poverty levels are higher; risk factors such as malnutrition are more prevalent; and resources are more limited.

In this paper, we go beyond the standard approach of evaluating an existing program, such as the work on Head Start (Bitler, Hoynes, and Domina, 2014; Kline, and Walters, 2016) and Early Head Start (Love et al., 2005). Instead, we design and evaluate with a clustered Randomized Controlled Trial (c-RCT) a *scalable* intervention aimed at improving an existing

[1] Cunha, Heckman, Lochner, and Masterov, 2006; Heckman, 2006; Engle et al., 2007; Doyle, Harmon, Heckman, and Tremblay, 2009; Almond, and Currie, 2011, Pongcharoen et al., 2012; Shonkoff, and Garner, 2012; Yoshikawa et al., 2013.

parenting program run by the government. The intervention we study involved the introduction to FAMI, *an existing government program,* of (i) *a structured early stimulation curriculum*, delivered through weekly group sessions with mothers and children, and monthly individual home visits; (ii) *training and coaching of the personnel delivering the intervention*, provided by trained mentors (tutors, henceforth); and (iii) an enhanced *nutritional supplement* for beneficiary children, alongside with nutrition education.[2] By collaborating with the government and using the existing infrastructure (i.e., program structure and personnel), we place the intervention within an operating institutional setting, which facilitates reaching scale.

The main question we are asking is whether offering early stimulation and appropriate nutrition in poor environments in a manner designed to be scalable by building on a nationwide program implemented by a government agency, can still improve child human capital and ultimately mitigate the effects of poverty. In our context, scalability of an intervention depends on its cost, but also on the possibility of running the intervention within an institutional framework that can handle it effectively. This is a key policy question, as well as one that adds to the evidence on the importance of early childhood interventions.

FAMI brings together mothers and their infants in a group setting with other mother-child dyads. Sessions are run by a local woman employed by the government, the FAMI mother. We developed a program adapted to these circumstances and inspired by the original Jamaica home visiting intervention (Grantham-McGregor, Powel, Walker, and Himes, 1991), now known as *Reach Up* (RU, see Grantham-McGregor, and Walker, 2015; Walker, Chang, Smith, Baker-Henningham, and the Reach Up team, 2018), and its replication in a scalable fashion in Colombia (Attanasio et al., 2014, Attanasio, Cattan, Fitzsimons, Meghir, and Rubio-Codina, 2020; Andrew et al., 2018).

The intervention was randomly allocated to 46 of 87 municipalities located in three of Colombia's 32 departments and lasted for an average of 10.4 months. Mothers in the control communities still had the option of attending the existing program (FAMI). In other words, the counterfactual against which treatment is compared is FAMI running as usual, and not the complete absence of the program (see Kline, and Walters, 2016).

On an intention to treat basis, our intervention significantly improved children's cognitive development by 0.16 (p-value 0.044) of a standard deviation (SD), with an implied average treatment on the treated (ToT) effect of 0.3 SD to 0.4 SD, depending on how we define compliance and intensity of treatment. As our end-line data were collected so that children were exposed to the treatment for at most 10 months, we also perform a dosage analysis, where variations in exposure were due to differences in the timing of the training of facilitators. Our analysis shows that the impact increases with increased intervention exposure. We also find some evidence of heterogenous impacts, with impacts larger for beneficiaries in the poorest households. This is consistent with the findings from another at-scale early years intervention (Bitler et al., 2014), although this study focuses on children older than those we consider.

[2] Moran, Ghate and van der Merwe (2004) and Nowak and Heinrichs (2008) reported that the most effective parenting programs included an evidence-based curriculum, systematic training of frontline workers, and opportunities for parents to learn and practice with children.

Children's nutritional status also improved: the fraction of children whose height-for-age is below -1 SD declined by 0.058 (p-value 0.098) with a corresponding increase in those with height-for-age between -1 SD and 1 SD (0.068 SD increase in height-for-age, p-value=0.046).[3] Results on the long-term effects of nutritional interventions are scarce and generally mixed. While short-term positive impacts were sustained in Guatemala (Hoddinott et al. 2013), in the Jamaica experiment, powdered milk supplementation showed important impacts early on which faded out in the longer term (Walker, Chang, Powel, and Grantham-McGregor, 2005; Walker, Chang, Powel, Simonoff, and Grantham-McGregor, 2006; Walker, Chang, Vera-Hernandez, and Grantham-McGregor, 2011). In both studies children were stunted at baseline. In Attanasio et al. (2014), micronutrient supplementation to a population of children with no specific nutritional deficit had no effect.

In addition to the main impacts, we also explore the mechanisms through which these might have been achieved. After providing evidence that the intervention significantly increased some potential mediators, such as parental investments in children, we show that indeed parental investment can explain most of the observed impact on child development using simple mediation analysis. This finding is confirmed by the results from a structural model that accounts for the endogeneity of parental investment in the estimation—a result consistent with that in Attanasio et al. (2020).

Our findings demonstrate the potential for improving human capital in poor settings and therefore form the basis for policy in a broader set of contexts across LMICs and contribute to the limited existing literature on the scalability of ECD interventions. The evidence on the long-term impacts of parenting interventions is mainly from small efficacy trials.[4] However, the evidence on the short- and medium-term impacts of *scalable* or *at-scale* parent support programs—that is, interventions designed to improve outcomes for a large number of children—is scarce and inconclusive both in high-income countries and LMICs.

Relevant studies in high-income countries include Robling et al. (2016) for the evaluation of NFP in the UK, Cattan, Conti, Farquharson, and Ginja (2019) for that of Sure Start in the UK, Love et al. (2005) for Early Head Start in the US, and Hjort, Solvsten, and Wüst (2017) in Denmark. In LMICs, most of the few existing studies report on short-term impacts—such as the evaluation of the nationwide Cuna Mas Program in Peru (Araujo et al., 2019), the evaluation of a group-based intervention delivered within a nationwide conditional cash transfer (CCT) program in Mexico (Fernald et al., 2017a), program integrations within primary health clinics in the Caribbean (Chang et al., 2015) and Bangladesh (Hamadani et al., 2019), or an evaluation comparing home visits vs group delivery in India (Grantham-McGregor et al., 2020). Two exceptions, which investigate impacts approximately two years after the end of intervention activities, are the studies in Colombia, where early stimulation and supplementation were

[3] The p-values we report are adjusted for multiple testing as explained in the main body of the paper.

[4] Examples from the US include the Nurse-Family Partnership (NFP) (Olds, Henderson, Tatelbaum, and Chamberlin, 1986a, Olds, Henderson, and Tatelbaum, 1986b; Olds, Henderson, and Kitzman, 1994; Heckman, Holland, Makino, Pinto, and Rosales-Rueda, 2017) and the Promising Practices (Brooks-Gunn et al., 1994; McCormick et al., 2006). In LIMCs, there is the well-known Jamaica home visiting model, which provided early stimulation (play-based activities) and nutritional supplementation (powdered milk) to stunted children in slums in Kingston for 24 months and obtained large impacts on ECD outcomes in the short term that translated into improved IQ and mental health (Walker et al., 2011) and higher wages (Gertler et al., 2014) in adulthood.

delivered within the infrastructure of the country's CCT (Attanasio et al., 2014; Andrew et al., 2018); and in Pakistan, where these were integrated into an existing community-based health service (Yousafzai, Rasheed, Rizvi, Armstrong, and Bhutta, 2014; Yousafzai et al., 2016). Scalability of effective and sustainable interventions is therefore a major and salient challenge.

The evidence we present also has direct implications for the importance of safety-net programs, such as Food Stamps in the US (see Hoynes, Schanzenbach, and Almond, 2016), for child outcomes. These programs can improve nutrition for children by providing more resources to parents. We show that providing such nutritional supplementation directly (in combination with child stimulation) can be an effective way of improving children's nutritional status, implying that parents do not appear to crowd out the additional resources provided for the children, even when they are delivered for use at home, as in our case. The absence of (complete) crowding out is a key element for understanding whether such programs can work and the extent to which they do.

The rest of the paper is organized as follows. The next section describes the context, the existing program, and the add-on intervention we evaluate. In Section 3, we discuss the evaluation design and sample. Section 4 presents the empirical strategy and section 5 the main evaluation results. Section 6 investigates the mechanisms behind the impacts obtained and, finally, section 7 discusses policy implications and concludes.

2. Background and intervention

The intervention that we evaluate consists in improving FAMI, an existing program run by the Colombian Family Welfare Agency (*ICBF* for its acronym in Spanish), a government institution. The fact that the innovation we are considering is grafted on a pre-existing infrastructure is important both for interpreting the size of its impacts and to provide a genuinely scalable model. In this section, we first describe the existing program and then describe the improvement that we test.

2.1 Description of the existing parent support program, FAMI

The FAMI program is aimed at supporting vulnerable families during pregnancy, childbirth and early childhood with nutrition, health monitoring and childrearing. Beneficiaries are identified by their score in SISBEN, Colombia's proxy means test based on household socio-economic characteristics and used for targeting most social policies. For the child stimulation component, the program is delivered through weekly group sessions of one hour each, and a monthly home visit of about an hour for parents of children 0-24 months of age. Group meetings take place in community spaces such as schools and churches, or the FAMI facilitator's own home. Based on ICBF's nationwide administrative data from 2013, prior to the beginning of this study, the size of each FAMI unit varies between 10 and 24 beneficiaries with a mean of 13 (SD=1.4). Approximately 80% of the beneficiaries are parents of children 0-24 months of age and 20% are pregnant women. Close to 225,000 families were FAMI beneficiaries around 2013 when this study started. FAMI mothers, the program facilitators, are local women and generally have a high school degree but no specific training on ECD. Similarly, the program has no concrete curriculum, other than some general operational guidelines and broad learning standards.[5] Indeed during the pilot stage we observed a rather diverse set of activities and discussions during the group sessions, with little-to-no engagement of the children. The monthly home visits were not designed around stimulation activities for the child but involved general advice to the family. The program also delivers a nutritional supplement that corresponds to 22% to 27% of the (monthly) recommended calorie intake of children younger than two and pregnant women. The average cost of the pre-existing FAMI program is $318 US (US dollars or USD) per child per year (Bernal, 2013). Further details on the pre-existing program and on the nature of the changes we introduced are provided in Appendix 1.

2.2 Description of the intervention

The intervention we evaluate aims to enhance the existing program through three complementary elements: (i) a structured early stimulation curriculum to improve child development, accompanied by pedagogical materials such as books, puzzles and toys; (ii) training and coaching for the FAMI mothers; and (iii) a larger and higher quality nutritional supplement than that previously received by FAMI participants, along with nutrition education during group sessions and home visits, and other materials such as recipe books and cards with age appropriate nutrition messages.

The stimulation curriculum was based on RU (Grantham-McGregor, and Walker, 2015; Walker et al., 2018), adapted, for the most part, to group meetings. FAMI includes, however, a monthly home visit, whose content was, again, adapted from RU. Both group meetings and home visits, last for about an hour, and aim at improving parenting practices and at introducing developmentally appropriate activities for children—in particular, activities that promote language, cognitive and fine motor development.

Mothers are encouraged to practice stimulation activities on a daily basis. Although most of the program content was delivered through the weekly group sessions, the monthly home visits were used to better tailor the activities to the developmental level of each child, and to introduce other, possibly more complex, activities. With respect to RU, the adapted curriculum added group discussions, more language activities, activities for children aged birth to 6 months, and cards with nutrition information. The program also trained mothers in sensitive and responsive parenting and appropriate behavior management, in promoting positive interactions, discouraging child mistreatment and ultimately promoting child socio-emotional development. The curriculum was designed to be delivered by facilitators without specialized knowledge of child development. For this reason, it was purposefully quite prescriptive.

Separate group meetings were offered for pregnant and lactating women with children up to 6 months, mothers with children 6 to 11 months, and mothers with children aged 1 to 2 years. However, as in practice mothers did not keep to their allocated slots, we ensured that the session would cater to children of different ages, with age-appropriate activities for all. An average of 5 mothers attended each session (min=1, max=15, SD=2.6). The curriculum involved

[5] This approach applies to all public ECD services in the country to date. The Board for Early Childhood has emphasized the principle of curricular freedom, and national standards are intentionally broad. Program providers are expected to adapt the learning standards to their own programs.

materials to be used during the sessions, including age-appropriate books, puzzles, home-made toys, pictures, construction blocks and nutrition cards. The intervention also included supplementary sessions to teach mothers how to construct home-made toys with recyclable materials that could be used to practice the activities proposed at home. This way, most mothers were able to set up a toy library for home use. All materials used in the session were taken home for practice and returned the following week.[6]

Pregnant women were invited to participate in all sessions and were encouraged to practice the activities along with the other mothers and their babies. However, in this study, we focus on the impacts of the intervention on children 0-24 months only.

A team of nine tutors, with college degrees in psychology and social work, trained and supervised by the research team, trained the FAMI mothers in the intervention before it started. Training was provided sequentially by town. All FAMI mothers in each given town were trained simultaneously for an average of 3.5 weeks and 85 hours.[7] The training involved demonstration, practice, and feedback in running the group sessions and in conducting the play and language activities with mothers and children, and in learning how to make the home-made toys. After the initial training was finalized, the tutors coached the FAMI mothers continuously throughout the duration of the intervention. In each supervision round, which took place approximately every 6 weeks, tutors observed one group session and one home visit, after which they provided feedback to the FAMI mother. Each tutor oversaw 5 towns and 19 FAMI mothers, on average. The tutors were, in turn, supervised by a program supervisor (a member of the research team) who visited each tutor every 2 months.

In short, the curriculum we introduced was intended to add both structure and content to the on-going sessions. FAMI mothers in the treatment group found the intervention to be substantially different to what was going on in the status quo, with 82% reporting they found it differed from their usual practice.[8]

Lastly, the intervention also included a monthly nutritional supplement which provided 35% of the daily calorie intake requirements for target children.[9] The nutritional content of the supplement was specifically targeted either for the pregnant mothers or to each child depending on their age—see Appendix 1 for further details.[10] All supplements were delivered monthly to the FAMI facilitator, who was in charge of distributing them among program participants during the first group session of each month. Families would not receive the monthly nutritional

[6] While we received authorization from the ICBF to implement and evaluate the intervention, its deployment was not publicized.
[7] This was done in two stages: an initial stage of 2 weeks and a second stage of 1.5 weeks about two months later, on average. More specifically, towns with less than 5 FAMI units received 75 hours of training in 3 weeks, towns with 6 to 9 FAMI units were trained for 100-125 hours in 5-6 weeks and towns with more than 10 FAMI units received training during 150-175 hours offered during 6-7 weeks.
[8] Specific differences with respect to how they had typically worked were: (i) practicing play activities with mothers and their children; (ii) practicing language activities with babies; (iii) making home-made toys with mothers; (iv) encouraging parents to play with their children at home; and (v) listening to parents about their achievements at home. Almost all of them (99%) reported that they would continue to use the proposed curriculum after the end of the project.
[9] In fact, it included more than that as it allowed for a potential consumption of up to 20% of the supplement's nutritional content by other household members.
[10] The package contained tuna, sardines, canola oil, iron-fortified whole milk (the only micronutrient included), beans, and lentils.

supplement if they did not attend this session. So, in a way, the early stimulation component represented a conditionality to receiving the supplement.

Clearly, crowding out of other nutrition and sharing within the household is a central concern. Participants were told that the beneficiary of the supplement was the child. However, there was no way to guarantee that its content was appropriately used in the home nor the extent to which it was (exclusively) offered to the target child.[11] We can only provide suggestive evidence, based on the program outcomes.

Table 1: Costs of the original program and its improvement
US $ per child per year

	Original program	Additional Intervention costs
Materials	$8	$27
Other administration costs	$2	-
Salary FAMI mother	$240	-
Mentoring	0	$88
Total without nutrition	**$250**	**$115**
Nutrition	$77	$209
Total with nutrition	**$327**	**$322**
FAMI training	N.A.	$11 one-time cost

Table 1 presents the running cost of the existing program, in the first column, alongside the additional cost of the intervention—the improvement we evaluate—in the second column. Costs are presented by component, showing a total program cost with and without nutrition. All values in Table 1 are expressed in USD per year per child, using the exchange rate at the time of the intervention, and assuming an average FAMI size of 10 mother-child pairs.[12]

The cost of the intervention we are evaluating, which is relevant both for its scalability and its cost-effectiveness, should not income the cost of the original program. As shown in Table 1, a substantial part of the cost of the original program is the salary of the FAMI mothers, which does not change, as the intervention did not hire additional FAMI mothers or decreased the number of children served by each FAMI. However, a substantial component of the cost of improving the existing program is the monitoring and mentoring that the FAMI mothers now receive. This amounts to $88 US per year per child, which covers the salaries of the tutors. For comparison, the FAMI mother salary corresponds to $240 US per child per year. Including the $27 US for materials yields a total cost of the coaching component of $115 US. Excluding the nutritional component in both the original program and this intervention, the FAMI intervention we are considering increases the cost of the program by about 46%. We consider the initial facilitator training ($11 US) as a one-off expense to be incurred in the first year. As it could benefit subsequent cohorts of children, it should be seen as an investment with some

[11] We could not evaluate the stimulation component alone—i.e., without the nutritional component—because both were part of the original program. Dissociating them for evaluation purposes was not feasible both logistically and ethically.

[12] This is conservative, given an average of 9.5 children younger than two per FAMI unit in the sample (plus 2.1 pregnant women); and a nationwide average of 13 (SD=1.4; range=[10, 24]), as computed using administrative data for 2013 (before the intervention started).

durability.[13] The largest increase in cost comes from the added nutritional package, which costs 2.71 times more than what it regularly costs—from $77 US to $209 US per child per year. Overall, the total increase in the cost of the program is of $322 US (or $333 US adding the one-off initial training), which effectively amounts to doubling the original cost of $327 US per child per year. Appendix 1 offers additional details on the cost of each component; and Appendix 2 includes a more thorough discussion on costs and scalability.

3. Sampling design, descriptive statistics, and implementation

The study took place between September 2014 and July 2016. At the start of the project, we prepared a pre-analysis plan and registered the trial at the ISRCTN registry (Appendix 8).[14] The intervention was intended to operate for 15 months between the end of 2014 and March 2016. In practice, the total duration varied by community, mainly to accommodate the initial training, and lasted an average of 45 weeks (10.4 months) with a range of 34-58 weeks. The logistics of rolling out the intervention implied a considerable amount of variation in exposure for the target children, mainly due to organizational issues.

The study towns were located in three departments in central Colombia (Cundinamarca, Boyacá and Santander). They were all chosen to have (i) fewer than 40,000 inhabitants, to avoid large urban centers; (ii) at least two FAMI units;[15] and (iii) no more than one unit of another public parenting program called *Modalidad Familiar* (MF) to minimize attrition towards this alternative program. MF is a public parenting program, similar to FAMI that was introduced during the first half of 2014.[16] The presence of MF is balanced between control and treatment sample towns, so that we are de facto estimating the effect of enhancing the FAMI program in the presence of some MF. Importantly for interpreting the results of our evaluation, the presence of MF in the study sample is minimal, with only 7% of the target children leaving FAMI to join MF. We further discuss this issue below.

Out of a universe of 135 such towns in these departments, we randomly drew 49 for the treatment group and 47 for the control. We assigned the remaining 39 towns to a randomly ordered waiting list. Towns in this waiting list were used to replace towns that had completely transitioned to the new MF program (whether in treatment or control). We could successfully replace 10 of the 19 towns that no longer ran the FAMI program, which yielded a final sample of 87 towns: 46 in the treatment group and 41 in the control group.

The average number of children younger than two per FAMI unit in the sample was 9.5 (SD=2.9) and the average number of pregnant women was 2.1 (SD=1.7). This implies an

---

[13] Whilst a similar argument on durability could be made for the materials, experience has taught us that their depreciation rate is quite high, as they are rotated among families. Hence, it is safe to assume that they do need to be replaced, approximately, on a yearly basis.

[14] The trial registration is at http://www.isrctn.com/ISRCTN93757590.

[15] This requirement is associated with the power calculations for the trial, and to facilitate the logistics associated with the training and coaching carried out by the tutors, who had to travel across various towns.

[16] MF is similar to FAMI in that it serves beneficiaries through monthly home visits and weekly group meetings but: (1) it serves children 0-5 years of age while FAMI serves children aged 0-2; (2) it has a set-up infrastructure for group meetings (a center) while FAMI uses other community spaces or the FAMI's own home; (3) serves, on average, 45 beneficiaries as compared to close to 15 in FAMI; (4) is led by a professional and an assistant, as compared to a single person who is not required to have a college degree in FAMI; (5) offers a nutritional supplement five times larger than that of FAMI; and (6) has access to a group of professionals including a psychologist and a nutritionist who support MF activities.

average of 11.6 (SD=2.8) total beneficiaries per FAMI unit. Within each unit, we enrolled in the study all children under 12 months of age at baseline, leading to a sample of *N*=1,460 children (4.3 children per FAMI and 17 per town, on average). We chose this subsample of children in order to maximize the potential time of exposure to our intervention, before children outgrew the FAMI program at age two. Overall, a total of 702 children in 171 FAMI units in 46 towns received the treatment (our enhanced version of the FAMI program); and 758 children in 169 FAMI units in 41 towns were in the control group, and therefore continued to receive the FAMI program as usual. At follow-up, we tried to reach all children in the study sample, regardless of whether they were still attending a FAMI or not, and regardless of the length of their exposure to FAMI.

Appendix 3 provides further details on the study design including power calculations, the study flow of participants, and the geographic distribution of treatment and control towns.

3.1 Data

As described in the pre-analysis plan, reported in Appendix 8, we defined a number of primary outcomes. These included measures of nutritional status—namely, externally standardized height-for-age Z-scores, constructed following the World Health Organization (WHO) standards (World Health Organization, 2006; World Health Organization, 2007); cognitive, receptive and expressive language, and fine and gross motor development, measured by the Bayley Scales of Infant and Toddler Development, third edition (Bayley, 2006); and socio-emotional development, as measured by the Ages and Stages Questionnaire: Socio-Emotional (ASQ:SE) (Squires, Bricker, and Twombly, 2009a). We chose developmental tests that have been extensively used in evaluations of early care or education and/or have been recommended for LMICs (Fernald, Pardo, Raikes, and Kariger, 2017b). These instruments were either available in Spanish, or had been previously translated, as had been used in Colombia before among similar populations. Anthropometric measures were collected in both rounds, whereas developmental measures were only collected at follow up. At baseline, children were younger than one year of age. Given the limited resources we had and how complex and expensive it is to reliably assess the development of such young children, we decided not to. [17]

For the analyses, we used internally age-standardized Bayley-III scores, where raw scores were standardized using the sample mean and SD calculated from weighted local smoothing regressions. We also aggregated all Bayley-III subscales using the factor model described in Appendix 4, which we interpret to reflect the child's 'cognitive' development. Children with extreme values for developmental or nutritional outcomes, according to international standards, were excluded from the analyses.[18]

In order to obtain an understanding of the mechanisms at play we also estimate impacts on intermediate outcomes that could have mediated the effect of the intervention on children's

[17] Child development assessments and anthropometric measures were collected by testers with degrees in psychology and health, respectively. The remaining variables in the household survey were collected by regular enumerators, prior to the child assessments.
[18] Specifically, we excluded 12 children who scored more than 3 SD below the mean on the Bayley-III cognitive scale (possible disability) and 15 children who were 6 SDs below the mean and 6 SD above the mean of height-for-age (extreme observations).

developmental outcomes. In particular, we collected by maternal report, both at baseline and at follow up, information on variables that measure the quality of the home environment, maternal self-efficacy, maternal knowledge about child development, and food insecurity.

For the quality of the home environment, we used four variables constructed from items in UNICEF's Family Care Indicators (FCI, Kariger et al., 2012)—the number of magazines, books, or newspapers in the home; the number of toy sources; the number of varieties of play materials in the home; and the number of varieties of play activities the child engaged in with an adult over the three days before the interview—which were summarized in a single factor, labelled 'parental investment' and estimated using the factor model described in Appendix 4. We assessed maternal self-efficacy using the self-efficacy in the nurturing role scale in Porter and Hsu (2003). This scale contains 16 items rated on 7-point scales that pertain to mothers' perceptions of their competence on basic skills required in caring for an infant. To measure maternal knowledge about child development, we used 10-items, some selected from the Knowledge of Infant Development Inventory (KIDI, MacPhee, 1981) and some developed by the research team.

Food insecurity was collected with the Latin American Scale for the Measurement of Food Insecurity (ELCSA scale), both at baseline and at follow up. The ELCSA had been previously validated in Colombia (ELCSA Scientific Committee, 2012) and allows classifying households in four food insecurity levels: secure, mild insecurity, moderate insecurity and severe insecurity (Colombia. Presidencia de la República, 2008). In the analysis, we use an indicator which equals 1 if the household is food insecure (mild, moderate or severe) and 0 otherwise.

Detailed socio-economic household information was also collected, including maternal vocabulary scores—a proxy for maternal IQ—which was assessed on the Spanish version of the Peabody Picture Vocabulary (PPVT or TVIP, Dunn, L.M., Padilla, Lugo, and Dunn, L.M., 1986).

Finally, background information on FAMI mothers was gathered directly from them in both rounds. In addition to basic socio-demographic characteristics, we also collected their vocabulary scores and knowledge on child development using the same tests as for mothers.

3.2 Descriptive statistics

Table 2 shows baseline characteristics by treatment status. At baseline, children were, on average, for both the treatment and control groups, 5.6 months of age and in about 27% of the cases the father was absent from their household. Households had two children, on average; maternal average schooling was 8.6 years; and 23% of mothers were teenagers. In 2010, the teenage pregnancy rate was 21% nationwide and 30% for young girls living in households in the poorest income quintile.

The target population was particularly poor: average household income was COP 501,000 per month (US 178) which represents 81% of the legal monthly minimum wage in 2014. Close to 70% of these households had answered the SISBEN survey for screening of social program eligibility—a good proxy for poverty—and 96% of those surveyed were deemed eligible for social programs (i.e., they scored in SISBEN levels 1 and 2). Similarly, 62% of households in the sample had a total income below the poverty line adjusted for household size. In 2014, the poverty rate was 42% in semi-urban and rural areas of Colombia.

Table 2. Sociodemographic characteristics of children and their families at baseline

	Treatment	Control	p-value	RW
Sociodemographic characteristics				
Child's age in months	5.72	5.51	0.353	0.976
	(3.39)	(3.26)		
Child's birth weight (gr)	3189	3156	0.442	0.981
	(572)	(500)		
Maternal age (number of years)	26.16	26.47	0.421	0.981
	(6.84)	(6.70)		
Maternal years of schooling	8.85	8.41	0.121	0.751
	(3.42)	(3.31)		
Household Income (COP thousands)	526.1	477.2	0.232	0.930
	(388.1)	(340.7)		
Household size	4.08	4.10	0.931	0.990
	(1.47)	(1.43)		
Maternal PPVT (raw score)	22.32	19.76	0.037	0.379
	(8.53)	(8.08)		
Child's gender (% male)	51.9	50.9	0.729	0.990
First born (%)	46.6	45.1	0.648	0.990
Teenage mothers (%)	25.4	20.9	0.059	0.567
Father present (%)	69.7	75.1	0.035	0.379
Owns home (%)	37.1	39.6	0.623	0.990
Household in poverty (%) [a]	58.7	64	0.298	0.950
Intermediate outcomes				
Parental Investment [b]	-0.03	0.03	0.625	0.866
	(0.96)	(1.02)		
Maternal knowledge [c]	29.26	29.49	0.680	0.944
	(3.61)	(3.44)		
Maternal self-efficacy	26.50	26.49	0.974	0.977
	(5.51)	(4.67)		
Food insecurity (%)	50.4	41.9	0.222	0.520
No. of observations	700	756		

Note: ***p<0.01; **p<0.05; *p<0.1. Standard deviations (clustered by town) in parentheses. RW: p-values adjusted for multiple testing using the Romano-Wolf (Romano, and Wolf, 2005; Romano, and Wolf, 2016) step-down method. In this case all hypotheses in the Table are included in the RW p-value calculation. a % of households with total income below the poverty line in 2014 ($50 US person/month). b Factor score of FCI subscales. c Only available at follow-up (raw scores presented).

The environment in which the sample children grew up is highly deprived: in terms of the home learning environment ('parental investment'), on average, these households owned 2.6 books, magazines or newspapers and 1.4 different varieties of play materials for young children in the household; and adults reported to have engaged in 2.5 different types of play activities with young children over the past 3 days.[19] For comparison, among a representative sample of low-middle-income households with children 6-12 months in Bogota (Colombia's capital city), we observed an average of 3.2 different varieties of play materials and 3.4 different types of play activities. Moreover, the median household in this sample only owned 3 books for adults.

[19] These variables are not shown in Table 1 but correspond to the components of the FCI 'parental investment' factor.

Table 3. Nutritional status of children at baseline by randomization status

	Treatment	Control	p-value	RW
Weight-for-age z-score	0.26	0.27	0.921	0.988
	(1.39)	(1.42)		
Length/height-for-age z-score	-0.01	-0.21	0.241	0.856
	(1.68)	(1.74)		
Weight-for-length z-score	0.37	0.55	0.167	0.829
	(1.59)	(1.65)		
Underweight (%)	6	5.1	0.423	0.936
Risk of underweight (%)	9	10.7	0.377	0.936
Wasting (%)	5.9	6.4	0.746	0.988
Risk of wasting (%)	10.6	8.2	0.159	0.829
Stunting (%)	9.2	13.9	0.075	0.574
Risk of stunting (%)	14.7	15.5	0.791	0.988
Overweight (%)	9.9	9.2	0.691	0.988
Obesity (%)	4.8	7.3	0.165	0.829

Note: ***p<0.01; **p<0.05; *p<0.1. Standard deviations (clustered by town) in parenthesis. Adjusted p-values using the Romano-Wolf (Romano, and Wolf, 2005; Romano, and Wolf, 2016) procedure (2,000 iterations, clustered by town) are included in the last column. All variables in the table are considered as one group of hypotheses. Underweight: weight-for-age < -2 SD; risk of underweight: weight-for-age between -1 and -2 SD; wasting: weight-for-height < -2 SD; risk of wasting: weight-for-height between -1 and -2 SD; stunting: height-for-age < -2 SD; risk of stunting: height-for-age between -1 and -2 SD; overweight: weight-for-height between 2 and 3 SD; obesity: weight-for-height > 3 SD.

In Table 3, we show averages for the baseline nutritional status of children by treatment status. Specifically, we report weight-for-age, height-for-age, and height-for-weight Z-scores, in addition to a variety of nutritional indicators by deficit or excess as identified by international standards. 12% of the children in our sample are stunted. For comparison, stunting was about 9.3% for children younger than one year of age in rural areas in Colombia in 2013 and 11.8% in urban areas (as measured in the Colombian Longitudinal Household Survey, ELCA, 2013). Table 3 also shows that an additional 15% of children were at risk of stunting—i.e., children whose height-for-age was between -2 SD and -1 SD.

Table 4. Developmental outcomes of children in the control group at follow-up

	Mean (Sd. Dev.)	N
Bayley		
Cognitive Composite Score	91.98 (13.07)	703
Language Composite Score	91.59 (12.31)	702
Motor Composite Score	93.97 (12.58)	701
ASQ:SE		
% of children at socio-emotional risk	0.23	705

Note: ***p<0.01; **p<0.05; *p<0.1. Standard errors clustered by town in parenthesis.
Bayley-III composites computed based on external standardization provided by test developers. The fraction of children at socio-emotional risk by the ASQ:SE is computed using the thresholds provided by the test developers (Squires, Bricker, and Twombly, 2009b).

In Table 4, we report the mean and standard deviation of the cognitive, language and socio-emotional development levels for the control group as measured at follow up (ages 17 to 33 months). These have been standardized with mean 100 and standard deviation 15, which is the US reference population (composite scores). Subject to all the caveats of such comparisons, this allows us to place our population relative to the expected developmental outcome under favorable conditions. The Bayley-III composite scores were 0.6 SD below the norming sample mean in both the cognitive and language scales, and 0.4 SD below in the motor scale. We also observed that 18% of children score between -1 SD and -2 SD with respect to the norming sample in cognition, 23% in language and 15% in motor development. Only about 2% to 3% would be considered at risk of developmental delay given that their composite scores are below -2 SD.

In terms of socio-emotional development, 23% of the children were at risk of developmental delay according to thresholds defined by the ASQ:SE using the test norming sample. For comparison, we know from the ELCA (2013) that 22% of children younger than two in low SES urban households were at risk of developmental delay by the same measure, 26% in high SES urban households and 19% in rural households in 2013.

Finally, in Appendix 5 we present basic characteristics of FAMI mothers by study group. On average, they were 42 years of age, had completed 13 years of education, and they had almost 12 years of work experience in the FAMI program. They had an average of 2.5 children of their own. There were no jointly significant differences between FAMI mothers in treatment and control towns.

3.3 Attrition, Compliance and Dosage

In both treatment and control towns, children in the sample might be 'lost' in the follow-up survey and/or might drop out of FAMI. The first is an attrition problem, while the latter is a compliance one. At follow up, we attempted to reassess all children, including those who dropped out of FAMI, to avoid non-random selection.

We report figures on attrition in Appendix 6 (Table 6.1). The attrition rate, of 8.6%, was slightly higher in the treatment group (10.5%) than in the control group (6.7%), although the difference is significant only at the 10% level. Children lost at follow-up were older, less likely to have a resident father at home, and more likely to have mothers with lower vocabulary (PPVT) scores. Moreover, as shown by the interactions of the treatment indicator with observables, attrition affected slightly the composition of the treatment and control samples (third column of Table 6.1). While the attrition differential between treatment and control towns was not very large, in Appendix 7, we discuss how we deal with the potential bias that it could introduce to our impact estimates. Furthermore, there we show that attrition does not bias our main findings.

Children who dropped out of the FAMI program between baseline and follow-up, if found, were interviewed at follow-up and their families were asked for the reason to leave FAMI. 47% reported that they outgrew the program eligibility age, 40% that they started attending a different ECD public program (12% a parenting program and 28% a childcare program), and 13% reported to have moved to another municipality. In Tables 6.2 and 6.3 in Appendix 6, we show that the treatment slightly reduced the probability of dropping out of FAMI for an alternative program and is not related to the probability of attending MF.

If age-eligible, a family could have attended a maximum of 44 weekly group sessions and received 11 monthly home visits during the study period. In terms of effective attendance, 77.5% of all children in the treatment group assessed at follow-up participated in at least one FAMI pedagogical activity (group session or home visit), while the rest did not attend any at all. Information on participation to specific activities was collected as part of the supervision protocol of the enhanced intervention and therefore is only available for the intervention group. In Figure 6.1 in Appendix 6 (graphs a. and b.), we show the distribution of children in the intervention group by total exposure to the pedagogical component of the program. Conditional on having attended at least one session, the median number of pedagogical activities attended was 28 out of a total of 55.[20]

On the main reasons why parents found it difficult to attend group sessions or receive home visits, close to 38% reported child illness, 15% reported maternal illness, and 19% reported conflict with other commitments. An additional 12% reported difficulties in finding or being able to afford transportation to the meetings and 10% reported bad weather. The remainder reported other reasons. Children with lower program attendance were older, less likely to live with their fathers, and had younger and more educated mothers. Whilst they exhibited better learning environments at home, they were exposed to higher verbal or physical punishment (Table 6.4 in Appendix 6).

Regarding, the nutritional component of the intervention, close to 29% of children in the treatment group did not receive any nutritional supplements and those who received at least one, received 9.8 supplements on average (SD=3.6) out of a maximum of 14 (Appendix 6, Figure 6.1, graph c). As the supplements were delivered by the FAMI mother during the first group meeting of each month, non-attendance implied that a beneficiary might not receive the supplement. We cannot verify if and how the nutritional supplement was used at home or the extent to which it was shared within the family.

Compliance with both components of the program largely overlapped the same subsamples of children. In particular, 69% of children in the treatment group received at least one nutritional supplement and attended at least one session, 19% did not receive any nutritional supplements nor attended any sessions, 8% attended at least one session but did not receive any nutritional supplements, and 4% received at least one supplement but never attended sessions (Figure 6.1, graph d. in Appendix 6).

4. Estimating average impacts

For each outcome of interest, we estimate *Intent to Treat* (ITT) effects on children's development using the regression:

$$Y_{isl,1} = \beta_o + \beta_1 T_{sl} + \delta' X_{isl,0} + F_{l,0}\sigma + D_0\theta + Z_{isl,1}\rho + \varepsilon_{isl,1} \qquad (1)$$

where $Y_{isl,1}$ is an outcome of interest for child i in FAMI unit s in town l at follow-up (t=1); T_{sl} is a dummy equal to 1 if the FAMI unit s in town l was in the treatment sample. $X_{isl,0}$ is a set of baseline child and household characteristics, including child's age, gender, and weight-for-

[20] Some other children in the treatment group might have dropped out of the FAMI program between baseline and the beginning of the intervention, due to the time elapsed to complete the training of the FAMI mother (up to four months). These children, therefore, would not have attended any of the enhanced sessions.

age and height-for-age z-scores, the household's wealth index, maternal PPVT scores (to proxy for maternal IQ) and an indicator for the mother being an adolescent. These are included to improve efficiency and to correct for any minor baseline imbalances caused by attrition.[21] Finally, D_0 represents a set of department fixed effects, which control for regional differences, $Z_{isl,1}$ is the vector of tester or interviewer dummies, and $\varepsilon_{isl,1}$ is the residual term. We cluster standard errors of the estimates at the town level, which is the unit of randomization.

The presence of the MF program in the town does not bias our impact estimates. MF was in place before randomization and our sample of children was drawn from those attending the FAMI center at baseline before randomization. Moreover, as documented in Appendix 6, treatment did not affect the probability of switching to MF and it only affected that of switching to other alternatives marginally.

In addition to average impacts, we look at impacts across the distribution of outcomes and also analyze the possibility of heterogeneous impacts in two ways. First, we consider the entire distribution of the outcomes of interest in the treatment and control samples and test for differences in these distributions using the Anderson-Darling statistics (Anderson, and Darling, 1952).[22] Second, we re-estimate equation (1) for subgroups in the evaluation sample. In particular, we divide the sample by wealth, as measured by a household wealth index, by mother's education and by child's gender.

5. The impact of the improved FAMI

For most outcomes, we measure impacts in terms of SD units of the variable of interest in the control group. We also include the 95% confidence interval, the standard p-value for two-tailed null hypotheses and the Romano-Wolf stepdown p-values adjusted for multiple hypotheses testing for the specific group of hypotheses presented in each table. The Romano-Wolf procedure was performed using 2,500 bootstrap replications and clustering by town.

5.1 Main impacts

In Table 5, we report the average impacts of the intervention on the Bayley-III factor for a summary measure of overall development; the ASQ:SE for socio-emotional development; and the height-for-age Z-score for nutritional status. In subsequent Tables, we present results for more disaggregated measures of these outcomes. Impacts are computed, regardless of whether children actually attended the program or how many times they attended, that is, these are OLS estimates of equation (1) or ITT.

The effect of the program on the Bayley-III factor was 0.163 SD and it is statistically significant at the 5%, after adjusting for multiple hypotheses testing for the three primary outcomes in the table. We find no significant average impact of the program on socio-emotional development or height-for-age Z-scores. Socio-emotional development is part of the set of

[21] Item non-response in baseline covariates is not correlated with treatment status. Thus, we imputed missing covariate values with the average of the non-missing observations and accounted for this imputation with a dummy variable in equation (1). The exact fraction of imputed observations varies by covariate up to a maximum of 6.8%.
[22] Such a test is considered more powerful to detect differences in the tails of the distribution than the Kolmogorov-Smirnoff test (Engmann, and Cousineau, 2011).

potential outcome variables as the program also aimed at training mothers in sensitive and responsive parenting and appropriate behavior management. However, the curriculum had a stronger focus on cognition and language through the demonstration and practice of specific activities, which might explain the lack of effect on socio-emotional development.[23] We discuss further the results on nutritional status below.

Table 5. Impact on children's outcomes

VARIABLE	Impact (95% CI)	P Value	RW P Value
Bayley-III Factor	0.163$^{(++)}$ (0.035,0.290)	0.015	0.044
ASQ:SE Total Score	0.021 (-0.096,0.139)	0.722	0.699
Height for age Z-Score	0.078 (-0.038,0.195)	0.190	0.335

Note: $^{(++)}$ $p<0.05$ based on Romano-Wolf adjusted p-values (RW, Romano, and Wolf, 2005; Romano, and Wolf, 2016), as we consider 3 simultaneous hypotheses for children outcomes. 95% confidence interval in parenthesis for two-tailed tests. Standard errors clustered by town. Covariates included: child's gender, household wealth index, maternal PPVT score, teenage mother and baseline weight-for-age and height-for-age Z-scores. Bayley-III factor is a factor score of the five age-standardized Bayley-III scales. ASQ:SE total score is the age-standardized ASQ:SE score.

As mentioned, the impacts in Table 5 are measured in terms of SD of the outcome of interest in the control group. An alternative meaningful metric would be the fraction of the gap in the outcome of interest that the estimated impact represents in a reference population. To perform such an exercise, we use a subsample of children analyzed by Rubio-Codina, Attanasio, Meghir, Varela, and Grantham-McGregor (2015). The authors considered a sample of about 1,400 children aged 6 to 36 months living in families representative of the bottom 85% of the wealth distribution in Bogota and estimated a difference in the Bayley-III cognitive scale of about 0.8 SD between those in the top and the bottom 25% of such wealth distribution, which correspond roughly to the 17th and the 68th percentile of the entire population in the city. To make the Bogota and the FAMI samples comparable, we estimated a factor model using both samples simultaneously, but limiting the Bogota sample to children of the same age as the FAMI children. We used the Bayley-III cognitive scale, available in both samples, as an anchor and imposed a loading factor normalized to 1. We find that the developmental levels of FAMI children are similar to those of children in the bottom 10% of the Bogota sample and the impact of the intervention equivalent to closing the gap between children in the top and bottom wealth decile by 23%.

The size of these effects is not negligible, especially if we take into account that the intervention lasted on average no more than 45 weeks and attendance was incomplete (77.5% attended *at least one* session). It also compares favorably to the impacts of nearly 0.26 SD obtained in Attanasio et al. (2014), which was a one-on-one weekly home visiting program that lasted for 18 months with very high compliance rates.

The role of attrition. As discussed earlier, there has been some attrition, which is differential between the treatment and control groups, even conditional on observables. To

[23] Note also that the measures used to capture socio-emotional development might not be very precise.

assess the possible bias caused by this, we estimate a selection model where attrition is a function of baseline characteristics as well as indicators for the identity of the interviewers assigned to households at baseline and follow-up. The identity of the interviewers explains attrition presumably because of differing quality among them. Furthermore, as interviewers were allocated randomly across towns making their identity orthogonal to individual characteristics, their identity is a valid instrument. We also need to assume that the identity of the interviewers is unrelated to children's outcomes, which is reasonable since those administering the Bayley-III test were different people to the interviewers collecting the household survey. The attrition equation is estimated jointly with the outcome equation. The results are reported in Table 7.1 in Appendix 7 and show that our conclusions are not sensitive to correcting for such non-random attrition.

5.2 Treatment on the Treated and dosage effects.

Treatment on the Treated effects. Since non-compliance with the program is one sided, we can use instrumental variables to identify the effect of Treatment on the Treated (ToT), using the random assignment to treatment as an instrument. There are, however, many different ways of thinking of the intensity of the program. If we measure effective participation as the fraction of children who attended at least one of the pedagogical activities of the program (i.e., a group session or a home visit), which is 77.5%, then the ToT on the Bayley-III factor is 0.21 SD. If, instead, we measure effective participation as the fraction of children in the treatment group who attended at least the unconditional median number of sessions (i.e., 21 out of 55 total), which is 53.2%, the ToT on the Bayley-III factor is 0.30 SD. Finally, if we define effective participation as the fraction of children who attended the median number of pedagogical activities conditional on having attended at least one (i.e., 28 sessions), which is 38.6%, then the ToT effect is 0.42 SD.[24] Thus, the potential effects are large even for a reasonably short intervention, delivered in groups. To realize such potential compliance, we would need to improve our understanding of the factors that drive attendance and whether parents misperceive the returns of the program in terms of child development. This is a key area of further research.

Dosage effects. By the time follow-up data were collected, the FAMI intervention had been running for about 10 months. This short interval was dictated by budgetary considerations. As discussed in section 2, the intervention involved training the FAMI mothers for about 3.5 weeks. The trainers, divided in several groups, covered all treatment towns in about 2 months. The end-line data collection itself extended for about two months. The combination of these two factors meant that by the time the outcomes were measured the potential intervention dosage that children could be exposed to in the various treatment communities varied considerably, between 34 and 58 weeks. We define potential dosage of the intervention as the number of sessions that could have been attended during the period comprised between the date

[24] There is an additional complication in estimating ToT effects from the ITT impacts we report. As we mentioned above, our estimate represents the impact of the improved FAMI relative to the standard FAMI (status quo), which is attended by the children in the control group. Presumably, there are also compliance problems in the control program on which, unfortunately, we do not have data. The ToT estimate we have discussed should be interpreted as the impact of a fully compliant improved FAMI over the business-as-usual FAMI in which compliance does not change.

in which the children were assessed at end-line and the date at which the training had been completed, divided by 10. For the control sample, dosage is fixed at 0. As this measure of dosage was determined by logistical considerations, it is very likely to be uncorrelated with child development outcomes and thus we assumed it is exogenous.

To corroborate this assumption, we test whether dosage correlates with a number of village variables within the treatment group. The results do not show any discernible correlation (see Table 6.4 in Appendix 6). Furthermore, we add to the observable controls in equation (1) a variable that measures the difference in days between follow up and baseline data collection rounds. This difference was also driven by similar logistic considerations but does not correlate with our measure of dosage.

Given this evidence, we modify equation (1) in the following fashion:

$$Y_{isl,1} = \beta_o + \beta_1 Dos_{sl} + \delta' X_{isl,0} + F_{l,0}\sigma + D_0\theta + Z_{isl,1}\rho + \varepsilon_{isl,1} \qquad (1')$$

where Dos_{sl} is dosage as defined as above. We report the results on the Bayley-III factor as the outcome of interest in Table 6.

Table 6. Effects of potential dosage on Bayley-III factor

VARIABLE	**Potential Dosage (St. Error)**	**Effect of Average Potential Dosage (p-value)**
Bayley-III Factor	0.209** (0.079)	0.169** (0.010)

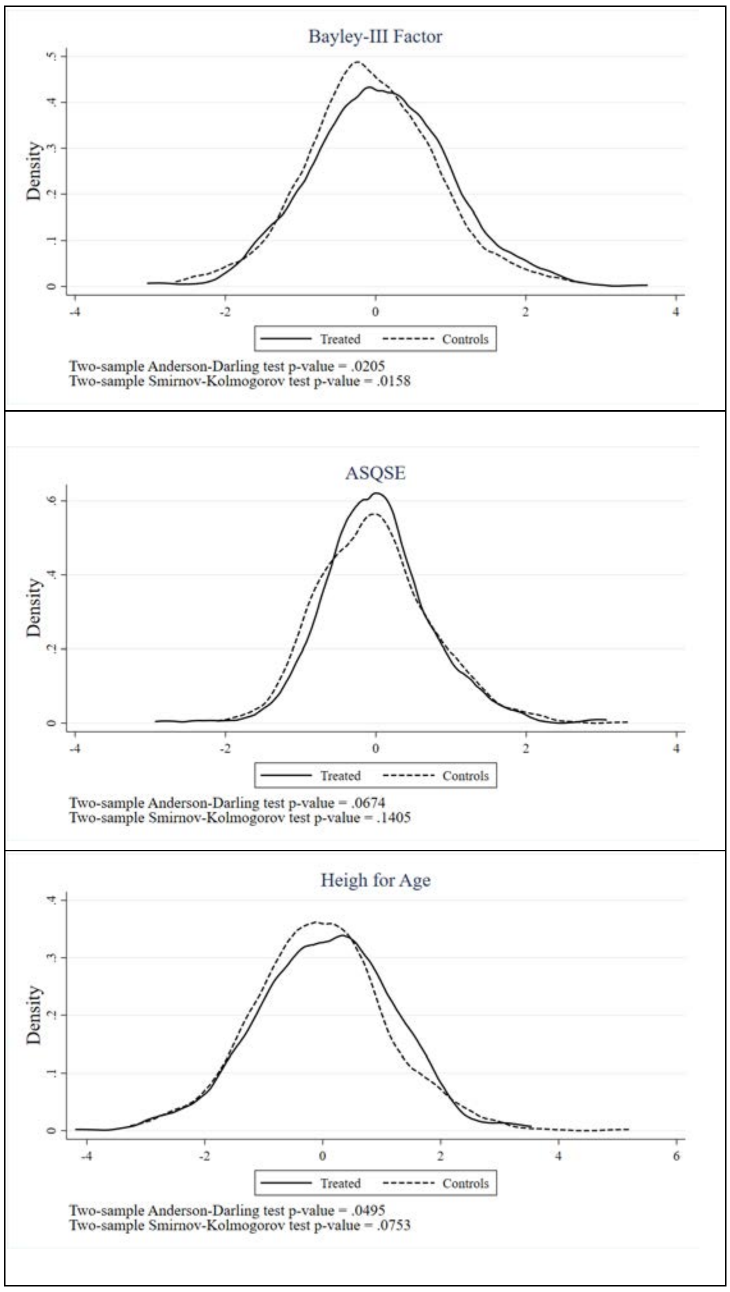
Note: ***p<0.01; **p<0.05. Standard errors clustered by town in parenthesis. Covariates included: child's gender, household wealth index, maternal PPVT score, teenage mother and baseline weight-for-age and height-for-age Z-scores, the difference in days between baseline and follow up data collections. In the treatment group the potential dose varies from 34-58 weeks.

The estimates show a positive and significant effect (with a p-value of 0.010) of dosage equivalent to an increase of 0.209 SD in cognitive development for every 10 additional sessions. In the last column of the Table, we report the impact implied by these results for the average dosage received by children in the treatment group, which is estimated at 0.169. This result is consistent with the impact reported in Table 5. We also experimented with a quadratic specification for dosage. We do not find any significant non-linearity. This result is perhaps not surprising given the relatively short amount of time the intervention had been implemented at the time we collected follow-up data.

5.3 Heterogeneous impacts

In this subsection, we look at heterogeneity in impacts. As mentioned in Section 4, we consider both unobserved heterogeneity and heterogenous impacts by observable variables, such as wealth and maternal education.

Figure 1. Distribution of conditional outcomes by treatment status



Note: Plot of the distribution of the residuals resulting from a regression of outcomes on observed characteristics described in equation (1), for the treatment and the control samples separately.

Unobserved heterogeneity. Figure 1 reports the distribution of the Bayley-III factor and the ASQ:SE (socio-emotional skills) by treatment and control. To obtain each figure, we first regress the respective outcome on the control variables included in equation (1) and then we plot the distribution of the residuals of this regression for the treatment and the control groups separately. In the graph, we also report the p-value of the Anderson-Darling (AD) and the Kolmogorov-Smirnov (KS) tests for the null hypothesis of identical distributions by groups.[25]

What is apparent from the graphs and the results of these tests is that the program had a significant impact on the Bayley-III factor (p-values =0.010 and 0.12 for the AD and KS tests, respectively) and affected the distribution over most of its support. The results for the ASQ:SE are less strong; nevertheless, the p-value for the AD test is 0.067, showing some impact.

As we saw in the descriptive analysis, 12% of the children in our sample are stunted (height-for-age < -2 SD) and 15% are at the risk of stunting (-2 SD < height-for-age < -1 SD).

[25] The Anderson-Darling test focuses more on the tails of the distribution and has been shown to have greater power than alternative tests, such as the Kolmogorov-Smirnov test (Bennet, 2008), which focuses on first order dominance.

It is well-established that stunting at this age is a good indicator of long-term malnutrition and can have long-run negative impacts on human capital development (Hoddinott et al., 2013). The program included a significant nutritional component, which, given the nature of our sample, could have both a short- and a long-term impact. While Table 5 did not show significant impacts on height-for-age, the third graph in Figure 1 shows a more nuanced picture and significant impact on the distribution of height for age (p-values=0.050 and 0.075).

We pursue this in Table 7, where we assess the impacts on different parts of the distribution of height-for-age. The results indicate that the fraction of children whose height-for-age was below -1 SD decreased by 5.8 percentage points or 0.13 SD, while the number of children with normal height-for-age increased by a similar fraction (6.8 percentage points). Both results are statistically significant at the 10% and 5% level respectively, even after adjusting the p-values for multiple testing and point to the value of considering the entire distribution. This result is of importance because it has often been proven difficult to impact height-for-age through less intensive interventions (s et al., 2019; Bernal, 2015).

Table 7. Impacts on height-for-age by ranges of the distribution

VARIABLE n1=597, n0=674	Impacts (95% CI)	p-value	RW p-value
Pr(Height-for-age between -5 SD and -1 SD)	-0.058$^{(+)}$ (-0.115,0.000)	0.054	0.098
Pr(Height-for-age between -1 SD and 1 SD)	0.068$^{(++)}$ (0.012,0.124)	0.020	0.046s
Pr(Height-for-age between 1 SD and 5 SD)	-0.011 (-0.035,0.014)	0.399	0.385

Note: $^{(++)}$ p<0.05 and $^{(+)}$ p<0.1 based on RW adjusted p-values (Romano, and Wolf, 2005; Romano, and Wolf, 2016), considering all three hypotheses jointly. Impacts measure the change in the probabilities considered in each row, in a linear probability model. Standard errors clustered by town. Covariates: child's gender, household wealth index, maternal PPVT score, teenage mother, and baseline weight-for-age and height-for-age Z-scores.

Observed heterogeneity. We now consider how average impacts differed across key groups. This exercise can help us understand whether the intervention helped the most vulnerable and from a policy perspective it helps improve targeting. We investigate whether the effects of the intervention on children's development, as measured by the Bayley-III factor, varied by maternal education, child gender, and household wealth at baseline.

For each of these three baseline variables, we divided the sample into two groups: less than high school versus more for maternal education; boy versus girl for child's gender; and household wealth above or below the sample median.[26] The results are reported in Table 8. Impacts do not seem to substantially vary by the level of maternal education. Although the point estimates are larger for mothers with complete high school (0.176 SD vs. 0.142 SD), this difference is not significant. Turning to gender, the point estimates suggest that the intervention worked better for girls, but the differences are, again, not significantly different from zero.

[26] The wealth index is computed as the first principal component of a number of dwelling characteristics (such as the material of walls, floors and roofs, the number of bathrooms and rooms, access to utilities, etc.) and durable goods ownership.

However, we do find significant effects of wealth on the impacts, even after correcting for multiple testing, across *all the six hypotheses* considered jointly. The effects, at 0.24 SD, are estimated to be much stronger for children living in poorer households. Moreover, the difference between the impact on children from poorer households and that on children from the higher wealth group is significant, with a RW p-value of 0.062.

Table 8. Heterogeneous impacts on the Bayley-III factor by child and household characteristics at baseline

Group (Number of observations)	Impacts (RW-p-value)	Estimated Difference (RW-p-value)
Maternal education ≥ complete high school (N=658)	0.176 (0.072)	0.034 (0.760)
Maternal education < complete high school (N=632)	0.142 (0.244)	
Male (N=671)	0.125 (0.077)	-0.074 (0.720)
Female (N=619)	0.198 (0.244)	
Wealth index above the median (N=655)	0.042 (0.599)	-0.243 (0.062)
Wealth index below the median (N=635)	0.285$^{(+++)}$ (0.006)	

$^{(+++)}$ p<0.01 based on RW stepdown p-values (Romano, and Wolf, 2005; Romano, and Wolf, 2016). Heterogeneous effects estimated by subsamples: Difference is a cross-model test for ITT associated parameter. Covariates included: child's gender (except in gender regression), household wealth index (binary), maternal PPVT score, teenage mother and baseline weight-for-age and height-for-age z-scores. Romano-Wolf stepdown p-values for the 6 multiple hypotheses for the impact and 3 hypotheses for the differences in the last column.

This result is key and contains both a positive and a negative message: the intervention can indeed improve the outcomes of the most deprived group in this already poor population. However, the better-off children from this group are in no way "well-off" or middle class and neither do they measure up well in their development against, say even the Bogota middle class, never mind the international standards. Hence the intervention would need to improve for this group. These results generally highlight the difficulty with improving ECD programs for broad populations—targeting interventions to the needs of separate groups is likely to be important. No significant heterogeneous effects were found in the case of socio-emotional or nutritional outcomes.

Lastly, we investigate whether intervention impacts varied by quality of implementation and FAMI mother characteristics. We do not find any significant difference on impacts by any of the measures of implementation fidelity available nor by FAMI mother's age or education. The only variable for which we find some marginally significant difference in impact is a measure of FAMI mother's 'motivation', as assessed by the tutors: children who attended centers by a FAMI mother reported to be more 'motivated' than the median, registered a higher impact (0.22SD vs 0.07). This 0.15 difference is significant with a Romano-Wolf p-value of 0.099, adjusted for 3 multiple hypotheses for process and fidelity of program implementation, and tutor-reported FAMI mothers' skills and motivation.

6. Understanding the impacts

In this section, we study possible mechanisms that could have generated the documented impacts on final outcomes. We start by estimating the impact of the intervention on a number of inputs that are relevant for child development, following Heckman, Pinto and Savelyev (2013). We then take a structural approach to estimate the causal link between the relevant inputs we consider and child development, taking into account the possible endogeneity of the former, through a production function framework similar to that in Cunha and Heckman (2008), Cunha, Heckman and Schennach (2010) and Attanasio et al. (2020).

6.1 Effects on intermediate outcomes and mediation analysis

The intervention we are studying is a transfer in kind of early education and nutritional supplementation. As with other transfers in kind, the intervention can induce parents to change their contributions to their child's development in other dimensions. The food supplement delivered by the intervention we are evaluating could be clawed back by reducing other food inputs to the target child, or perhaps sharing it in the family and even selling it; and the additional stimulation received by the target children could cause parents to switch attention to other children or to themselves, therefore mitigating the intervention's impact. On the other hand, it is also possible that low-income parents are not fully aware of the returns to investing in their children (Cunha, Elo, and Culhane, 2013; Attanasio, Cunha, and Jervis, 2019) so that the effects of the intervention may have been generated by an increase in investment induced by a change in these beliefs. Therefore, there are also good reasons to believe that, instead of crowding out, the intervention could have led to a crowding in of resources. In this case, adding to the transfer from the intervention may have particularly high returns. Indeed, Attanasio et al. (2020) evaluate another early years stimulation intervention in Colombia and show that, in response to it, parents crowd-in resources by increasing investments Exploring the mediating factors and the mechanisms underlying intervention impacts is a way of obtaining answers to some of these questions. Moreover, understanding these is critical to improve the design and targeting of public policies.

Table 9. Program impacts on intermediate outcomes

VARIABLE	Impact as fraction of SD in control group (95% CI)	P Value	RW P Value
Parental Investment	0.304 (0.207,0.472)	0.000	0.000
Maternal Knowledge (Raw Score)	-0.016 (-0.160,0.128)	0.831	0.836
Maternal Self-Efficacy (Raw Score)	0.039 (-0.108,0.186)	0.604	0.823
ELCSA Food Insecurity Status	-0.089 (-0.231,0.052)	0.169	0.399

Note: 95% confidence interval in parenthesis for two-tailed tests. OLS estimation; standard errors clustered by town; Impacts are measured in terms of SD of the control group. p-values are computed using Romano-Wolf (Romano, and Wolf, 2005; Romano, and Wolf, 2016) step-down procedure. We consider 4 hypotheses. Covariates Included: child's gender, household wealth index (binary), maternal PPVT score, teenage mother and baseline level of the outcome. Parental Investment is measured by a factor model estimated using the subscales of FCI Home Environment Quality, as discussed in Appendix 4.

We start by presenting, in Table 9, the effects of the program on the intermediate outcomes described in section 3.1. The first row reports the impact of the intervention on parental investment, estimated from the FCI index, which captures the quality of the home environment, and combines books, magazines and newspapers, play activities and play materials in the home (see Appendix 4). The following rows assess impacts on maternal knowledge about child development, maternal self-efficacy, and food insecurity. Maternal knowledge and self-efficacy as potential mediators capture the idea that, through the intervention, parents (mothers, in particular) might become more effective in their childrearing practices.

The impact on the quality of the home environment was 0.34 of a SD in the control group and statistically significant, with a p-value of zero. This is a strong result and indicates that the intervention induces parents to invest more in their children. However, we do not find any statistically significant program effects on maternal knowledge about child development, maternal self-efficacy or food insecurity.[27]

6.2 A structural interpretation of the impacts: production function estimates

Given the results on intermediate outcomes, we proceed to estimate a model where child development is determined by a production function which depends on parental investment and other background variables. Both child development and parental inputs are represented by latent variables which are not observed directly but for which we have informative markers that allow us to estimate them by factor analysis. Given the evidence in Table 10, the sole mediator we consider for child development is parental investment. This approach is a similar to that of Heckman et al. (2013). However, here, following Attanasio et al (2020), we also consider the possible endogeneity of parental investments.

We estimate a production function for human capital development, which we assume to be a function of parental investment, several other environmental factors and, potentially, the intervention itself. In particular, we assume that child development can be expressed by the Cobb Douglas production function:

$$lnCD_{isl} = \gamma_o + \gamma_1 \ln(PI_{isl}) + \gamma_2 T_{sl} + \delta' X_{isl} + F_l \sigma + D\theta + Z_{isl}\rho + u_{isl} \tag{2}$$

where CD_{isl} is the child development latent variable and PI_{isl} represents the parental investments latent variable, both estimated by the factor model described in Appendix 4 and used to estimate the reduced form impacts in Tables 5 and 10. In equation (2), the treatment allocation T_{sl} can affect child development both directly and through its impact on parental investments (*PI*). The covariates X_{isl} include the child's gender, household wealth, maternal PPVT score, a dummy variable for teenage mothers, and distance to the municipality's Town Hall to capture unobserved differences in household socio-economic condition. We also control for baseline childcare attendance and municipality population. Earlier studies also control for

[27] The effect we found is not as strong as that reported in Attanasio et al. (2014) of 0.5 SD on play materials and play activities with adults at home and resulting from a home visiting intervention in Colombia. We return to this issue in the Discussion section.

lagged child development. However, as explained, we did not collect baseline developmental outcomes since the children were too young to obtain a precise measure with the resources we had available. Instead, we control for the child's nutritional status at baseline—namely, height-for-age and weight-for-age. As before, D represents department fixed effects and Z_{isl} is the vector of tester fixed effects. Finally, u_{isl} represents unobservable factors determining child development, including shocks experienced by the child and additional inputs not observed by the researchers but possibly chosen by parents. The Cobb Douglas assumption is consistent with the evidence in Cunha et al. (2010) and in Attanasio et al (2020), who performed a similar analysis on another early stimulation intervention in Colombia delivered through home visits rather than group sessions.

The main challenge in estimating the parameters in equation (4) is the fact that parental investment PI_{isl} is likely to be endogenous, as the parents might be reacting to shocks experienced by the child or might choose investment jointly with other inputs. While the treatment is exogenous by construction—since it is assigned randomly across communities—it is not necessarily a valid exclusion restriction because it can have an independent effect on the outcome. Indeed, a question we pose is whether the treatment affects child development directly or whether its impact is mediated by parental investment. To answer this question, we need to establish the causal link from investment to child development. We therefore need an instrument, W_{isl}, that affects parental investment while not affecting child development directly. For this purpose, we use the travel time from the household residence to the FAMI center. To control for differences between households that are centrally located versus households that live in more outlying areas (that could differ in unobservable dimensions) we control for distance to the Town Hall when estimating equation (2) by instrumental variables (IV). Therefore, we estimate a first stage investment equation of the form:

$$\ln\,(PI_{isl}) = \pi_o + \pi_1 T_s + \pi_2 W_{isl} + \gamma' X_{is} + v_{is} \tag{3}$$

where the covariates X_{is} are the same as those in the production function in equation (2).

In the first column of Table 10, we report the treatment effect on the Bayley-III factor estimated by OLS; and in the second column, we introduce parental investment, also using OLS. The coefficient on treatment is reduced in size and it is no longer statistically different from zero, demonstrating that if the OLS assumption is valid, the impact is mediated by parental investments (although we cannot necessarily ignore the coefficient on treatment because it is quite large albeit imprecisely estimated).

In the third column of Table 10 we report the estimates of the investment equation coefficients π_1 associated to treatment allocation and π_2 associated to travel time to FAMI, which serves as an instrument when we estimate the production function shown in the subsequent columns. This is strongly significant, even conditional on distance to the Town Hall, which is intended to capture how centrally the household is located. Importantly the F-statistic is large enough to rule out a weak instrument problem, whether treatment is used an additional exclusion restriction or not (see bottom of column (3)).

Table 10. IV estimation of the production function for Bayley-III factor

	OLS Bayley-III factor (1)	OLS Bayley-III factor (2)	First Stage Parental investment (3)	IV Bayley-III factor (4)	 Bayley-III factor (5)
Treatment (*T*)	0.135**	0.079	0.294***	0.006	
	(0.065)	(0.065)	(0.068)	(0.110)	
Parental Investment (*PI*)		0.185***		0.467*	0.454***
		(0.036)		(0.249)	(0.171)
Time to Town Hall	-0.099***	-0.079***	-0.040	-0.048	-0.049
	(0.027)	(0.028)	(0.030)	(0.043)	(0.035)
Time to FAMI			-0.143***		
			(0.035)		
First stage F-statistics					
IV: Time to FAMI			16.86		
IV: Time to FAMI and Treatment			19.15		
Overidentification p-value					0.956
Observations	1292	1,292	1,292	1,292	1,292

Note: * p<0.01, ** p<0.05, * p<0.1. Standard errors clustered by town in parenthesis. Covariates included: child's gender, household wealth index (binary), maternal PPVT score, teenage mother and baseline weight-for-age and height-for-age z-scores, childcare attendance and town's population range, and interviewer and department FE.

In the fourth column of Table 10, we re-estimate the production function, as in column (2) but using IV. These estimates show a much higher impact of investments and a zero direct effect of treatment: the point estimates imply that the entire effect of treatment is driven by an increase in parental investments through the intervention. The difference of the investment coefficients in column (2) and (4) from 0.185 to 0.467 is significant at the 10% level and consistent with the results reported in Attanasio et al. (2020) where the coefficient in the production function of child development also increased considerably after accounting for the endogeneity of parental investment. This suggests that parents are compensating for negative shocks when choosing investment.

Given this last consideration, in the fifth column of Table 10, we remove the intervention from the production function (3). Now the coefficient on investment is 0.454 and it is significant at the 1% level. We notice that the model is now overidentified, as we now have two instruments for the single endogenous variable, PI_{is}. When testing the implied overidentifying restriction, we do not reject the null of correct specification.

7. Discussion and conclusions

Interventions that promote ECD, starting from birth, may well be the key to successful human capital policies, particularly in poor environments. However, the characteristics and the effectiveness of such programs at scale are not well understood, yet. In recent years, many early years interventions have been implemented worldwide, but effective and sustainable programs at scale are rare. Furthermore, many institutionalized initiatives are of low quality (Lo, Das, and Horton, 2017). Scaling up is not only a question of funds, but also of the available human resources in a variety of different contexts. A possible approach to deploying early years intervention at scale is to determine whether existing large-scale programs (and their infrastructure) can be successfully improved, so to guarantee the quality required for them to have significant impacts on children.

In this study, we present results from an experiment where we designed and implemented a *scalable* intervention that was added to an existing government group-based parenting support intervention, combined with nutritional supplementation. Effectively, the intervention we study is an improvement of an existing national program, consisting of incorporating structured content (curriculum of activities) and training and coaching for program facilitators, as well as nutrition education and a larger and higher quality nutritional supplement. As we have discussed, this design offers a directly scalable policy, both in terms of its costs and in its implementability, given the existing infrastructure and human resources.

Our curriculum is an adaptation of RU, a home visitation program shown to be effective in altering the long-run cognitive trajectory of children from deprived environments in its original implementation in Jamaica (Walker et al., 2011; Gertler et al., 2014). Adaptations of the curriculum to a variety of contexts and countries have also had positive impacts on developmental outcomes (see Grantham Mc-Gregor, and Smith (2016) for a review).

Evaluation of group-based adaptations of RU, and other parenting programs, are however more limited. Yet, they represent a promising and natural low-cost approach to improving outcomes in vulnerable populations in a more efficient manner as delivery is less intensive in human resources. Furthermore, while the delivery of the RU curriculum in groups might imply a reduced focus on the specific needs of an individual child, well-run groups might induce positive effects by improving existing networks and acquaintances and provide role models for some mothers.

The fact that we find reasonably-sized positive impacts in the short time span covered by our data collections is important—in practice, the interventions would last longer, and children would hopefully graduate into pre-schools where they could gradually build up their abilities and school readiness, thus addressing one key cause of poverty persistence. The evidence we present also points to potentially large gains where they are most needed, namely among the poorest. The importance of these results is even more apparent if we consider the fact that compliance with the number of sessions actually attended by children and their caregivers was relatively low and the intervention was relatively short, at least in comparison with the most successful efficacy trials referred to in this study. And yet our intervention had an ITT effect of 16% of a SD and a ToT effect of up to 42% of a SD in development. Moreover, there was a reduction in the fraction of children whose height-for-age is below -1 SD of 5.8 percentage points.

Some features of this particular study make us believe that these estimates are lower bounds of the potential of this intervention. First, the control group had access to the basic program, without the improved intervention—unlike similar studies in the literature in which the control group does not receive any intervention. Second, as stressed, the average impact reflects larger impacts for the children most in need and a small or null impact for the better-off children. Third, and most importantly, it was not possible to fully control and enforce the many relevant implementation aspects that might be needed to ensure fidelity of the intervention and impact development.[28] In fact, the implementation of the intervention was far from smooth and faced various challenges. Examples of the problems encountered included the low duration of participant exposure to the program, logistical difficulties for the delivery of pedagogical materials and the nutritional supplement in complicated rural geographies, heterogeneity in the fidelity of program implementation, and initial resistance of program providers to change their behavior. The implementation problems we document in our context are common to many programs implemented at scale.

The focus on the scalability is one of the most salient aspects of this study and reflects the difficulties policy makers face when moving from small trials to larger studies with reduced control over what actually happens in the field. As we suggest above, when an intervention is scaled-up, one needs to consider not only financial costs but also the possibility of sustaining quality of implementation given the existing service infrastructure. On the latter, we notice that our intervention was implemented on top an existing program, with a minimal involvement on the part of the researcher team. Our results indicate that, despite a number of implementation problems, which were in part present because we wanted to work with a model that could be reproduced at scale, the enhancement we evaluated had a sizeable effect on the children most in need. However, we do recognize that it is not obvious that a scaled-up intervention could maintain the level and quality of training and mentoring that were achieved during the study, although we stress that the evaluation did not use personnel with special qualifications. In any case, it is clear that proper mentoring should be developed with care.

Regarding the financial cost of the intervention, we notice that the cost of the pedagogical component of the intervention was US$115 per child per year (US$27 for pedagogical materials and US$88 for coaching) plus a US$11 one-off cost per child for FAMI pre-service training. At scale, there could be important economies of scale in the mentoring system, by far the largest component of the total pedagogical cost, which could reduce these figures substantially. The cost of the additional nutritional supplementation was US$209 per child per year. By the end of this study, the Colombian government adopted the nutritional supplementation evaluated herein nationwide, with an investment of US$10 million. The pedagogical component corresponds to 40% of the operational cost of the unenhanced version of the FAMI program, equivalent to 1.7 monthly minimum wages per year. In contrast, center-based childcare services cost US$1,100 per child per year. Or the transition to large childcare centers, which has been one of the center pieces of recent government policy, costs US$780 per child per year, more than twice the intervention we are studying. Therefore, the cost of our intervention is moderate, especially, in comparison to other ECD programs in the country, and financially sustainable.

As we stressed above, the impacts of the intervention we evaluated are relative to a status quo where children of the same age were receiving an unimproved program. To interpret these results, it is useful to put them in the context of the quality of other public early years services in Colombia. Bernal (2013) presents a diagnostic of public childcare quality by modality, using

[28] FAMI providers continued to be paid and supervised by the government with no legal obligation or additional monetary incentive to participate in our program. They were strongly encouraged to do so, but they could choose not to without any practical consequence.

standard measures. Quality levels are low for all modalities, close to minimum standards. This pattern is also found in other Latin-American countries. Part of the problem is precisely the lack of a structured curriculum and supervision/mentoring strategies—which is what the improvement we evaluate introduces to FAMI. What we show is that scaling up services with quality is possible within an existing institutional infrastructure that allows for such coaching and mentoring strategies. The evidence we presented suggests that it is possible to gradually improve the quality of nationwide programs at scale in a way that is affordable. Ours is an enhancement of an existing program that leverages on local low-skilled human resources. Critically, the intervention specifically aims at improving process quality (such as the integration of a structured curriculum and improved interactions between caregivers and children supported by coaching and mentoring), which the literature has shown to be critically associated with child developmental outcomes (NICHD, 2000; Yoshikawa et al., 2016)

A key question is whether these short-term impacts sustain over time. Andrew et al. (2018) report that the effects on child development and parental investment documented in Attanasio et al. (2014) disappear two years after the end of the intervention. The authors mention that this result might be due to a small initial effect (similar to ours) and/or the lack of continued family support for early stimulation. The impact fade-out observed for the intervention studied by Attanasio et al. (2014) is not unique. Several studies have found that medium-term program impacts might vanish but reappear later in the child's life-cycle (Schweinhart et al., 2005).

In Attanasio et al (2014), intervention activities ended as soon as the study ended. In our case, however, the intervention effectively kept running since an important part of it consisted of the training of the facilitators in the pre-existing program. In particular, most treated FAMI providers continued to use the curriculum although they were no longer being coached. In addition, participants in public programs are more likely to continue to be enrolled in similar public programs as children grow. For example, children could have moved on to home-based childcare, provided through the *Hogares Comunitarios* program (Bernal and Fernández, 2013), which could help reinforce or maintain these effects over time.

The total number of FAMI beneficiaries has decreased since 2013. However, still close to 150,000 children are still part of this program. Crucially, the toolkit developed for this intervention is flexible and easily adaptable to any ECD programs facilitated by paraprofessional personnel, as many are in Colombia, as well as in other developing countries. As we discuss in detail in Appendix 2, it would be straightforward to replicate *at scale* the training and coaching strategy proposed in this study by leveraging on the already existing monitoring and supervision infrastructure for community-based programs, including FAMI. Training professional staff in local ICBF offices would be feasible and they could easily implement both, training and coaching of FAMI and similar programs ran by paraprofessional personnel.

While the pre-existing program is present everywhere in Colombia, we implemented and evaluated the improvement in Central Colombia. This choice was motivated by the fact that this region tends to be more culturally and ethnically homogeneous with respect to other parts of Colombia, such as the coastal regions (both Pacific and Atlantic) where afro-Colombians and indigenous) are more likely to reside. Scale up in these regions would likely require additional piloting and adaptation.

To conclude, we show that a scalable program can have substantial effects on child development in highly deprived populations at a low cost and based on government infrastructure. Improving quality of large-scale programs in developing countries can form a key element of the policy toolkit for fighting poverty.

References

Almond, Douglas, and Janet Currie (2011). "Human capital development before age five." In *Handbook of Labor Economics*, Edited by O. Ashenfelter and D. Card (ed), Handbook of Labor Economics, edition 1, volume 4, chapter 15. Pages 1315-1486*,* Elsevier.

Anderson, Theodore W., and Donald A. Darling (1952). "Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes". *Annals of Mathematical Statistics*, 23(2), 193-212.

Anderson, Theodore W., and Herman Rubin (1956). "Statistical inference in factor analysis." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 5, edited by Jerzy Neyman, 111-50. Berkeley, CA: University of California Press.

Andrew, Alison, Orazio Attanasio, Raquel Bernal, Lina Cardona, Sonya Krutikova, and Marta Rubio-Codina (2019). "Preschool quality and child development." NBER Research Working Papers 26191, National Bureau of Economic Research, Inc.

Andrew, Alison, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina (2018). "Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomized controlled trial in Colombia." PLOS Medicine, 15(4), e1002556.

Araujo, María Caridad, Marta Dormal, Sally Grantham-McGregor, Fabiola Lazarte, Marta Rubio-Codina, and Norbert Schady (2019). "Home visiting at scale and child development." Unpublished manuscript.

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, Flavio Cunha, and Pamela Jervis (2019). "Subjective parental beliefs. Their measurement and role." NBER Research Working Papers 26516, National Bureau of Economic Research, Inc.

Attanasio, Orazio P., Camila Fernández, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina (2014). "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early childhood development program in Colombia: Clustered randomized controlled trial." *British Medical Journal*, 349, g5785.

Bayley, Nancy (2006) "Bayley scales of infant and toddler development." 3rd ed. Pearson Education.

Becker, Gary S. (1964). "Human Capital: A theorical and empirical analysis with special reference to education." National Bureau of Economic Research, edition 1, New York.

Bennett, Christopher (2008). "Consistent integral-type tests for stochastic dominance." Unpublished manuscript, Vanderbilt University.

Bernal, Raquel (2013). "The cost of early childhood policy in Colombia." Unpublished manuscript Universidad de los Andes.

Bernal, Raquel (2015). "The impact of a vocational education program for childcare providers on children's well-being." *Economics of Education Review*, 48, 165–183.

Bernal, Raquel, Orazio Attanasio, Ximena Peña, and Marcos Vera-Hernández (2019). "The effects of the transition from home-based community nurseries to childcare centers on children in Colombia." *Early Childhood Research Quarterly*, 77(2): 418-431.

Bernal, Raquel, and Camila Fernandez (2013). "Subsidized childcare and child development in Colombia: effects of Hogares Comunitarios de Bienestar as a function timing and length of exposure." *Social Sciences and Medicine*, 97, 241-249.

Bitler, Marianne P., Hilary W. Hoynes, and Thurston Domina (2014). "Experimental evidence on distributional effects of head start." NBER Working Papers 20434, National Bureau of Economic Research, Inc.

Black, Maureen M., Susan P. Walker, Lia C. H. Fernald, Christopher T. Andersen, Ann M. DiGirolamo, Chuling Lu, …, and Sally Grantham-McGregor (2017). "Early child development coming of age: science through the life-course." *The Lancet,* 389, 77–90.

Bradley, Robert H., and Diane L. Putnick (2012). "Housing quality and access to material and learning resources within the home environment in developing countries." *Child Development*, 83(1), 76-91.

Britto, Pia R., Stephen Lye, Kerrie Proulx, Aisha Yousafzai, Stephen G. Matthews, Tyler Vaivada, T., … and Early Childhood Development Interventions Review Group, for the Lancet Early Childhood Development Series Steering Committee (2017). "Nurturing care: promoting early childhood development." *The Lancet*, 389, 91-102.

Brooks-Gunn, Jeanne, Cecilia M. McCarton, Patrick H. Casey, Marie C. McCormick, Charles R. Bauer, Judy C. Bernbaum, …, and C. L. Meinert, (1994). "Early intervention in low-birth-weight premature infants: Results through age 5 years from the Infant Health and Development Program." *Journal of the American Medical Association,* 272(16), 1257-1262.

Cattan, Sarah, Gabriella Conti, Christine Farquharson, and Rita Ginja (2019). "The health effects of Sure Start." Working papers, The Institute for Fiscal Studies.

Chang, Susan M., Sally Grantham-McGregor, Christine A. Powell, Marcos Vera-Hernández, Florencia López-Bóo, Helen Baker-Henningham, and Susan P. Walker (2015). "Integrating a parenting intervention with routine primary health care: A cluster randomized trial." *Pediatrics*, 136(2), 272-280.

Colombia. Presidencia de la República, Agencia Presidencial para la Acción Social y la Cooperación Internacional, Instituto Colombiano de Bienestar Familiar, Departamento Adminstrativo Nacional de Estadística, Universidad de Antioquia, Programa Mundial de Alimentos, and FAO (2008). "Adaptación y validación interna y externa de la Escala Latinoamericana y Caribeña para la medición de seguridad alimentaria en el hogar - ELCSA-Colombia: Componente adaptación lingüística de ELCSA." Medellín.

Cunha, Flavio, Irma Elo, and Jennifer Culhane (2013). "Eliciting maternal expectations about the technology of cognitive skill formation." NBER Working papers 19144, National Beurau of Economic Research, Inc.

Cunha, Flavio, and James J. Heckman (2008). "Formulating, identifying and estimating the technology of cognitive and noncognitive skills formation." *Jornal of Human Resources*, 43(4), 738-82.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010). "Estimating the technology of cognitive and non-cognitive skill formation." *Econometrica*, 78 (3), 883–931.

Devercelli, Amanda E., Rebecca K. Sayre, and Amina D. Denboba (2016). "What do we know about early childhood development policies in low and middle income countries?" Washington, DC: World Bank Group. http://documents.worldbank.org/curated/en/252561473963612937/What-do-we-know-about-early-childhood-development-policies-in-low-and-middle-income-countries (accessed February, 2019).

Doyle, Orla, Colm Harmon, James. J. Heckman, and Richard E. Tremblay (2009). "Investing in early human development: timing and economic efficiency." *Economics & Human Biology*, 7(1), 1-6.

Dunn, Loyd M., Eligio R. Padilla, Delia E. Lugo, and Leotra M. Dunn. (1986). "Test de Vocabulario en Imágenes Peabody (TVIP)." MN: AGS Circle Pines.

ELCSA Scientific Committee (2012). "Escala Latinoamericana y Caribeña de Seguridad Alimentaria (ELCSA). Manual de uso y aplicación." Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), ISBN 978-92-5-307354-2.

Encuesta Longitudinal Colombiana de la Universidad de los Andes (ELCA). (2013). Retrieved from: https://encuestalongitudinal.uniandes.edu.co/es/datos-elca/2013-ronda-2

Engle, Patrice L., Maureen M. Black, Jere R. Behrman, Meena Cabral de Mello, Paul J. Gertler, and Lydia Kapiriri (2007). "Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world." *The Lancet*, 369(9557), 229–242.

Engmann, Sonja, and Denis Cousineau (2011) "Comparing distributions: The Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test." *Journal of Applied Quantitative Methods*, 6(3), 1-17.

Fernald, Lia C., Rose M. Kagawa, Heather A. Knauer, Lourdes Schnaas, Armando Garcia, and Linette M. Neufeld (2017a). "Promoting child development through group-based parent support within a cash transfer program: experimental effects on children's outcomes." *Developmental Psychology*, 53 (2): 222-36.

Fernald, Lia C. H., Elizabeth L. Prado, Abbie Raikes, and Patricia K. Kariger (2017b). "A toolkit for measuring early childhood development in low- and middle-income countries." Washington DC: World Bank.

Gertler, Paul, James J. Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, …, and Sally Grantham-McGregor (2014). "Labor market returns to an early childhood stimulation intervention in Jamaica." *Science*, 344(6187), 998-1001.

Grantham-McGregor, Sally, Akanksha Adya, Orazio Attanasio, Britta Augsburg, Jere Behrman, Bet Caeyers, …, and Karishma Vats (2020). "Group Sessions or Home Visits for Early Childhood Development in India: A Cluster RCT." *Pediatrics*, 146(6), e2020002725.

Grantham-McGregor, Sally, Christine A. Powell, Susan 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.

Grantham-McGregor, Sally, and Joanne A. Smith (2016). "Extending the Jamaican early childhood development intervention." *Journal of Applied Research on Children: Informing Policy for Children at Risk*, 7(2), 4.

Grantham-McGregor, Sally, and Susan Walker (2015). "The Jamaican early childhood home visiting intervention." *Early Childhood Matters*, 28-34.

Hamadani, Jena D., Syeda F. Mehrin, Fahmida Tofail, Mohammad I. Hassan, Syeed N. Huda, Helen Baker-Heningham, …, and Sally Gratham-McGregor (2019). "Integrating an early childhood development program into Bangladeshi primary health-care services: an open-label, cluster randomized controlled trial." *The Lancet*, 7(3), 366-375.

Heckman, James J. (2006). "Skill formation and the economics of investing in disadvantaged children." *Science*, 312(5782), 1900–1902.

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 papers 23610, National Bureau of Economic Research, Inc.

Heckman, James J., Rodrigo Pinto, and Peter Savelyev (2013). "Understanding the mechanisms through which an influential early childhood program boosted adult outcomes." *American Economic Review*, 103(6), 2052-86.

Hjort, Jonas, Mikkel Solvsten, and Miriam Wüst (2017). "Universal Investment in Infants and Long-Run Health: Evidence from Denmark's 1937 Home Visiting program." *American Economic Journal: Applied Economics*, 9 (4): 78-104.

Hoddinott, John, Jere R. Behrman, John A. Maluccio, Paul Melgar, Agnes R. Quisumbing, Manuel Ramirez-Zea, …, and Reynaldo Martorell (2013). "Adult consequences of growth failure in early childhood." The American Journal of Clinical Nutrition, 98(5), 1170–78.

Hoynes, Hilary, Diane W. Schanzenbach, and Douglas Almond (2016). "Long-run impacts of childhood access to the safety net." *American Economic Review*, 106(4), 903–934.

Jacoby, Hanan (2002). "Is there any intrahousehold 'Flypaper effect'? Evidence from a school feeding programme." *Economic Journal*, 112(476), 196-221.

Kariger, Patricia, Edward A. Frongillo, Patrice Engle, Pia M. Rebello Britto, Sara M. Sywulka, and Purnima Menon. (2012). "Indicators of family fare for development for use in multicountry surveys." *Journal of Health, Population and Nutrition*, 30(4), 472–86.

Kline, Patrick, and Christopher R. Walters (2016). "Evaluating public programs with close substitutes: the case of HeadStart." *The Quarterly Journal of Economics*, 131(4), 1795-1848.

Lo, Selina, Pamela Das, and Richard Horton (2017). "Early childhood development: the foundation of sustainable development. Comment." *The Lancet*, 389(10064), 8-9.

Love, John M., Ellen E. Kisker, Christine Ross, Helen Raikes, Jill Constantine, Kimberly Boller, ..., Cheri Vogel (2005). "The effectiveness of early head start for 3-year-old children and their parents: lessons for policy and programs." *Developmental Psychology*, 41(6), 885-901.

McComick, Marie C., Jeanne Brooks-Gunn, Sthephen L. Buka, Julie Goldman, Jennifer Yu, Mikhail Salganik, …, and Patrick H. Casey (2006). "Early Intervention in low-birth-weight premature infants: Results at 18 years of age for the Infant Health and Development Program." *Journal of Pediatrics*, 117(3), 771-780.

McPhee, David (1981). "Knowledge of infant development inventory." Chapel Hill, NC: University of North Carolina.

Moran, Patricia M., Deborah Ghate, and Amelia van der Merwe (2004). "What works in parenting support? A review of the international evidence." Policy Research Bureau Research, Report No. 574. Retrieved from https://dera.ioe.ac.uk/5024/1/RR574.pdf.

Neville, Helen, Eric Pakulak, and Courtney Stevens (2015). "Family-based training to improve cognitive outcomes for children from lower socioeconomic status backgrounds: Emergin themes and challenges." *Current Opinion in Behavioral Sciences*, 4, 166-170.

Nowak, Christoph, and Nina Heinrichs (2008). "A comprehensive meta-analysis of Triple P-Positive Parenting Program using hierarchical linear modeling: Effectiveness and moderating variables." *Clinical Child and Family Psychology Review*, 11(3), 114-144.

Olds, David, Charles R. Henderson, and Harriet Kitzman (1994). "Does prenatal and infancy nurse home visitation have enduring effects on qualities of parental caregiving and child health at 25 to 50 months of life?" *Pediatrics, 93*(1), 89-98.

Olds, Davis, Charles R. Henderson, and Robert Tatelbaum (1986b). "Preventing child abuse and neglect: a randomized trial of nurse home visitation." *Pediatrics*, 78(1), 65-78.

Olds, Davis, Charles R. Henderson, Robert Tatelbaum, and Robert Chamberlin (1986a). "Improving the delivery of prenatal care and outcomes of pregnancy: a randomized trial of nurse home visitation." *Pediatrics*, 77(1), 16-28.

Pongcharoen, Tippawan, Usha Ramakrishnan, Ann M. DiGirolamo, Pattanee Winichagoon, Rafael Flores, Jintana Singkhornard, and Reynaldo Martorell (2012). "Influence of prenatal and postnatal growth on intellectual functioning in school-aged children." Archives of Pediatrics and Adolescent Medicine, 166(5), 411-416.

Porter, Chris L., and Hui-Chin Hsu (2003). "First-time mothers' perceptions of efficacy during the transition to motherhood: Links to infant temperament." *Journal of Family Psychology,* 17(1), 54-64.

Robling, Michael, Marie-Jet Bekkers, Kerry Bell, Christopher C. Butler, Rebecca Cannings-John, Sue Channon, …, and David Torgerson (2016). "Effectiveness of a nurse-led intensive home-visitation programme for first-time teenage mothers (Building Blocks): A pragmatic randomised controlled trial." *The Lancet*, 387(10014), 146–155.

Romano Joseph P., and Michael Wolf (2005). "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73(4), 1237-1282.

Romano Joseph P., and Michael Wolf (2016). "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing." *Statistics & Probability Letters*, 113, 38-40.

Rubio-Codina, Marta, Orazio Attanasio, Costas Meghir, Natalia Varela, and Sally Grantham-McGregor (2015). "The socio-economic gradient of child development: cross-sectional evidence from children 6-42 months in Bogota." *The Journal of Human Resources*, 50(2), 464-483.

Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores (2005). "The High/Scope Perry Preschool study through age 40." Ypsilanti, MI: High/Scope Press, 194-215.

Shonkoff, Jack, and Andrew S. Garner (2012). "The lifelong effects of early childhood adversity and toxic stress." *Pediatrics*, 129(1), e232-e246.

Singla, Daisy R., Elias Kumbakumba, and Frances E. Aboud (2015). "Effects of a parenting intervention to address both maternal psychological wellbeing and child development and growth in rural Uganda: a community-based, cluster randomized trial." *Lancet Global Health*, 3(8), e458-469.

Squires, Jane, Diane Bricker, and Elizabeth Twombly (2009a). "Ages & Stages Questionnaires: A parent-completed child monitoring system." Baltimore, MD: Paul H. Brooks Publishing.

Squires, Jane, Diane Bricker, and Elizabeth Twombly (2009b). "Technical Report on ASQ:SE." Baltimore, MD: Paul H. Brookes Publishing.

Walker, Susan P., Susan M. Chang, Christine A. Powell, and Sally Grantham-McGregor (2005). "Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-retarded Jamaican children: a prospective cohort study." *The Lancet* , 366(9499), 1804-1807.

Walker, Susan P., Susan M. Chang, Christine A. Powell, Emily Simonoff, and Sally Grantham-McGregor (2006). "Effects of psychosocial stimulation and dietary supplementation in early childhood on psychosocial functioning in late adolescence: follow-up of randomised controlled trial." *BMJ*, 333:472.

Walker, Susan P., Susan M. Chang, Joanne Smith, Helen Baker-Henningham, and the Reach Up Team. (2018). "The Reach up early childhood parenting program: origins, content, and implementation." *Zero to Three*, 38(4), 37-43.

Walker, Susan P., Susan M. Chang, Marcos Vera-Hernandez, and Sally Grantham-McGregor (2011). "Early childhood stimulation benefits adult competence and reduces violent behaviour." *Pediatrics*, 127(5), 849-857.

World Health Organization (2006). "WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development." Geneva: World Health Organization.

World Health Organization (2007). "WHO Child Growth Standards: Head circumference-for-age, arm circumference-for-age, triceps skinfold-for-age and subscapular skinfold-for-age: Methods and development." Geneva: World Health Organization.

Yoshikawa, Hirokazu, Christina Weiland, Jeanne Brooks-Gunn, Margaret R. Burchinal, Linda M. Espinosa, Wiliam T. Gormley, …, and Martha J. Zaslow (2013). "Investing in our future: The evidence base on preschool education." New York, NY: Foundation for Child Development and Ann Arbor, MI: Society for Research in Child Development.

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When does preschool matter? *The Future of Children*, 26(2), 21-36.

Yousafzai, Aisha K., Jelena Obradovic, Muneera A. Rasheed, Arjumand Rizvi, Ximena A. Portilla, Nicole Tirado-Strayer, …, 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." *Lancet Global Health*, 4(8), e548–e558.

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.

Supplemental Materials.

Appendix 1. Detailed description of the interventions

Our program was aimed at improving the FAMI government program. During our initial formative research stage, we observed several group sessions in FAMI units located in municipalities not in our c-RCT sample. These were mostly lecture-like meetings in which the FAMI program facilitator would discuss a certain topic with beneficiary mothers sitting around a table and typically aided with a flipchart and/or other visual aids. Most importantly, babies would remain in their strollers or sitting on the mothers' laps with no direct participation in the sessions. The topics covered varied and some examples during our visits included community support, friendship, breastfeeding, dental care and teenage pregnancy, although sessions around sensitive parenting, child development and engagement of parents and other family members were also supposed to be addressed. In spite of such a wide spectrum of topics, however, the existing program did not use any curriculum: what to cover and how was entirely up to the FAMI facilitator.

The curriculum we introduced aims at assisting mothers to provide developmentally appropriate activities for their children (in particular, activities that promote language, cognitive, and motor development), as well as reinforcing maternal knowledge and practices about feeding and nutrition. In doing so, it aims at improving mothers' knowledge, practices, enjoyment in child up-bringing, and self-esteem. Given that the FAMI program is delivered mostly through group meetings and home visits, the intervention we designed included two complementary curricula. Both used similar components, actions, and activities to promote better maternal child rearing practices. These included making the mother the agent of change and empowering her to improve her child's development by demonstrating the use of age-appropriate play materials and activities, by providing opportunities to practice with them, and by providing supportive feedback. The program also aimed at training mothers in sensitive and responsive parenting and appropriate behavior management, and encouraged positive mother-child interactions and child maltreatment prevention.

Most of the program content was delivered through the group meetings as they were held on a weekly basis. In addition to being spaces where to demonstrate and practice the use of age-appropriate play materials and language activities, the groups provided opportunities for discussing and practicing effective child rearing skills and positive interactions with children with other caregivers, sharing experiences, group problem-solving, as well as opportunities for social support. Group meetings also provided the opportunity for mothers to discuss how play activities promoted children's development and show them how to make simple toys so that each family could set up a toy library for home use. Group meetings were one hour long. An average of 5 mothers attended each session (min=1, max=15, SD=2.6).

The home visits were delivered monthly and provided the opportunity to introduce activities that were more difficult (i.e., more specific to a child's developmental level) in the context of the group (such as puzzles and matching activities), additional language activities, and specific ideas on how to use routine home activities to promote child development and how to identify materials in the home that could be used to promote child development. Home visits were, on average, one hour long.

Separate group meetings were offered for pregnant and lactating women with children up to 6 months, mothers with children 6-11 months, and mothers with children aged 1-2 years, and mothers were asked to attend the meeting that corresponded to her child's age. However, in practice, this did not always occur, and, in anticipation, the curriculum had been designed so that it could be delivered to groups with children over the entire age range. To this end, the play and language activities were divided into age bands (birth-5 months, 6-11 months, and 1-2 years) by level of difficulty. We expected mothers of children 6-24 months to attend four meetings per month and pregnant and lactating women with children up to 6 months to attend two meetings per month.

Each group session was structured in six different moments: arrival and free play; feedback from the previous group session (10 minutes); song (5 minutes); demonstration and practice of the age-appropriate play and language activities for the week with materials that will be taken home (30 minutes); discussion around a parenting theme or activity (15 minutes); review of the session to ensure that mothers understand the activities, and commitment to practice with children at home (10 minutes); and in closing, they share a snack. The themes for discussion during the group meetings included issues such as the importance of spending time playing with the child, praising the child, talking to the child, things to do at bath time or mealtimes, learning to trust, understanding the child's feelings, teaching the baby about her environment, and child behavior.

Similarly, each home visit consisted of i) greeting and discussion of any issues, ii) feedback from the previous home visiting session, iii) song, iv) introduction of new play and language activities (including how to integrate into everyday routines, v) nutrition message, and vi) a review of activities to be conducted over the next month.

The curriculum included discussion topics or key parenting messages, age-appropriate activities to promote child development using the play materials, as well as everyday activities to encourage adult-child interactions. It was specifically designed to increase the focus on language development with respect to the original RU curriculum. For example, 1) the language activities were designed to be more structured and to involve demonstration and practice, in the same way as the other play activities, rather than being based on discussion, and 2) we included the use of materials in the language activities to make them more concrete (e.g. when encouraging mothers to label and talk about things in the home environment, we used relevant objects, such as a cup, comb, chair, during the session for demonstration and practice). It was very rich in play materials—books, pictures to talk about, home-made toys, puzzles and building blocks—which were used during home visits and in the group meeting. It also included a set of nutrition cards relevant to the children's ages that were discussed with the mother during each home visit. The complete kit of materials had a cost of $27 US per child per year.[29]

In addition to the set of activities and materials, the qualification of the FAMI program also included a training and coaching component (pre- and in-service training) to support and maintain the quality of home visits and group meetings. Shifting away from a supervision model, the new approach consisted of a team of tutors with degrees in psychology and social work, who provided the initial pre-service training and then continued to provide in-service training and support during the implementation period. Tutors, trained and supervised by the research team, were in charge of training FAMI mothers. Pre-service training was provided sequentially by town, where all FAMI mothers were trained simultaneously. Average training time was of 3.5 weeks and 85 hours, although total training time varied by town depending on the number of FAMI units. Specifically, towns with less than 5 FAMI units received 75 hours of training in 3 weeks; towns with 6 to 9 FAMI units received 100-125 hours for 5-6 weeks; and towns with more than 10 FAMI units received 150-175 hours of training over 6-7 weeks. The training involved demonstration, practice and feedback in running the group sessions and conducting the play and language activities with mothers and children, and toy-making. The

[29] All costs are computed at the average exchange rate of 2015-2016 (the evaluation period), of 2,800 COP/USD.

one-time cost of the pre-service training per FAMI provider was of about $113 US or $11 US per child.

Tutors also coached FAMI mothers continuously throughout the duration of the intervention. In each supervisory round, they observed one group session and one home visit, and provided feedback to the FAMI mother. Each tutor was in charge of 5 towns and 19 FAMI mothers, on average, and hence met with a FAMI mother every 6 weeks on average. Whenever possible, they also facilitated a group meeting of FAMI mothers in each town to discuss and share positive experiences and challenges and engage in problem-solving. The tutors were supervised by an intervention supervisor—the fieldwork manager, a member of the research team—who conducted visits with each tutor every 2 months.

Training costs included: tutors' salaries (plus fringe benefits), the fieldwork manager's salary, traveling expenses of tutors and supervisor, the cost of the tutors' training by the research team (accommodation, transportation, food, trainer's traveling expenses and accommodation, materials), and incentives offered to FAMI mothers for their participation ($14 US per FAMI). This was costed for 5 months, which was the total duration of training (for both, trainers and FAMI mothers). Intervention fixed costs, such as this one, were amortized over 171 treated FAMI units and assuming an average FAMI size of 10 children. Mentoring/coaching costs were similar to training costs and included: tutors' salaries, the fieldwork manager's salary, and travel expenses for both, tutors and supervisor. The difference is that mentoring is a recurrent cost for every month in which it took place. Overall, the cost of coaching was $82 US per month per FAMI provider or $8 US per child per month.

In addition to the introduction of the early stimulation curriculum, the intervention also included a nutritional component, which comprised the delivery of a monthly nutritional supplement to FAMI participants, and psychoeducation around feeding and nutrition during group meetings and home visits. The nutritional supplement corresponded to 35% of daily calorie intake requirements for pregnant women, breastfeeding mothers, and children younger than 2 years of age (for 30 days). The cost of the package is $26 US per month including shipping costs. It contains tuna, sardines, canola oil, iron-fortified whole milk, beans, and lentils. In terms of educational contents, we developed a cooking book that takes into account the socioeconomic characteristics of households in our sample, brochures on food-handling and classification, and 19 nutrition cards that were discussed with the mother during each home visit. Mothers received a nutrition card relevant to their child's age at these monthly home visits. The topics covered included things like breastfeeding, bottle-feeding, breastmilk extraction and storage, weaning, hygiene, finger foods, menu ideas, mealtimes, and chatting while feeding.

ble 1.1. Composition of the nutritional supplement

Nutritional supplement unenhanced FAMI					**Additional supplementation of the intervention**					**Total nutritional supplement**			
Calories	Protein	Calcium	Iron	Folic Acid	Calories	Protein	Calcium	Iron	Folic Acid	Calories	Protein	Calcium	Iro
16.1%	23.1%	33.8%	22.2%	56.6%	19.0%	31.1%	24.8%	11.0%	36.9%	35.1%	54.2%	58.5%	33.2
3.2%	4.1%	0.9%	7.1%	7.8%	8.0%	18.2%	20.1%	22.3%	42.4%	11.2%	22.2%	21.0%	29.4
7.6%	12.6%	1.5%	11.0%	17.2%	19.1%	56.2%	32.2%	34.7%	70.8%	26.6%	68.8%	33.7%	45.7

l intake requirements.

Appendix 2. The cost and scalability of FAMI in the Colombian context

2.1 Costs

The intervention we study costs about \$322 per year per child plus \$11 one-off cost per child for FAMI pre-service training. The cost of the unenhanced FAMI program is about \$327 per child per year. The pedagogical enhancement (excluding the nutritional supplement) corresponds to approximately 35% of the cost of the unenhanced version of the program. This is equivalent to 1.5 monthly minimum wages per child per year, or 2.5 monthly minimum wages per year including the nutritional supplement. For comparison, the cost per child per year in center-based childcare in Colombia is approximately \$1,100 or 4.4 monthly minimum wages per child per year.[30] In Table 2.1 we compare the costs and impacts of other interventions recently implemented in Colombia, with FAMI.

Bernal (2015) studies the impact of vocational training of the women running the family nurseries considered. She reports a sizeable impact at a low cost per child. Bernal, Attanasio, Peña, and Vera-Hernandez (2019) consider the transfer of children from home-based daycare services offered in the provider's own home to large childcare centers and find virtually no impacts at a very large cost.[31] Finally, Andrew et al. (2019) study the impacts of (1) targeted pedagogical improvements to center-based care in large cities and (2) staffing of these centers with nutritionists and psychologists. The impacts are comparable to ours at a slightly higher cost for the pedagogical component. Incidentally, the hiring of professional personnel in centers had no effects on children's cognition.

This summary highlights the importance of enhancements to what is known in the specialized literature as process quality (such as the integration of a structured curriculum and improved interactions between caregivers and children supported by coaching and mentoring) with respect to changes and improvements in the so-called structural quality alone (such as infrastructure, as in Bernal et al. (2019); or staffing, as in the second intervention studied in Andrew et al. (2019)). In particular, the former seems to have more cost-effective impacts with respect to the latter. In sum, the evidence we have presented shows that it is possible to gradually improve the quality of nationwide programs at scale in a way that is affordable, while maintaining quality and with a reasonably sized impact on children's developmental outcomes.

[30] Cost computations are not adjusted to consider the possibility of crowding in different programs since the intervention we evaluate did not affect the probability of leaving a FAMI to join the new program MF (although it did affect the probability of leaving a FAMI to join an alternative ECD service).
[31] The cost reported in the table corresponds to the difference in the cost per child/year in a childcare center and the cost in a family nursery.

Table 2.1. Costs and impacts of alternative quality enhancements of ECD programs in Colombia

Program	**Age of children at baseline**	**Cost child/yr** (USD)	**Impact** (SD of standardized scores)	**Duration of intervention** (months)	**Detailed result** Effect (SE or p-value)
Training paraprofessional care providers (Bernal, 2015)	6 months-5 years of age	101	0.25	12	+5.2 (2.65) score points for children 0-3 years of age on ASQ language scores, and +3.5 (2.0) score points for children 3+ years on Woodcock-Muñoz (WM) mathematical ability.
Transfer from home-based to center-based childcare (Bernal et al., 2019)	6 months-5 years of age	780	0.05	10-18	+0.05 (0.02) SD on nutrition factor, and -0.11 (0.05) SD on ASQ cognitive factor.
Targeted pedagogical improvements in center-based care (Andrew et al., 2019)	18-36 months of age	373	0.15	18	+0.15 (0.076) SD on cognition, language & school readiness (based on TVIP, WM cognition, Daberon and pencil tapping test)
Staffing of center-based care with professionals (Andrew et al., 2019)	18-36 months of age	150	0.10	18	Null effects on cognition. +0.1 (0.06) SD on height for age for children older than 30 months of age.
This study	0-12 months of age	320	0.15	10	+0.16 (p-value 0.044) on Bayley-III factor and -0.06 (p-value 0.54) reduction on the probability of being stunted or in risk of stunting.

2.2 Scalability

When thinking about the scalability of the FAMI improvement, we consider the possibility of using the already existing supervision infrastructure at ICBF. The national ICBF office works with 33 regional offices and 203 local (municipality-level) offices. Each local office has three different teams in charge of the main tasks. One of the teams, called the community team, is composed of 1 nutritionist, 2 pedagogues, 1 sociologist, 1 social worker and 1 educator. This team oversees supervision and monitoring of ECD services. The other two teams are responsible for child protection services and intra-household legal issues.

Supervision and monitoring of ECD providers is through regular on-site visits structured around a checklist. This checklist is based upon items that we would consider structural quality features, such as, the physical characteristics of the center, the cleanliness of bathrooms, furniture, etc. It does not cover process quality aspects related to the quantity and nature of the interactions in the classroom (Bernal, 2015). A bad evaluation may result in the closure of a center or the need to improve specific aspects to be re-checked within a certain period.

Based on our findings, we consider it would be extremely useful and effective to replace this type of supervision for one that focuses on process quality and uses a reflective approach that encourages best practices. As shown, this could be done by promoting the use of a structured curriculum and session guidelines that identify key features—both in content and form of delivery—that are relevant for child outcomes. These items could be assessed by supervisors, who could provide constructive feedback rather than penalties on the basis of their observations. In this way, continuous improvement would be supported.

There are close to 14,600 FAMI units countrywide. We think it would be possible to train 3 out of the 6 community team members (1 sociologists or 1 educator plus 2 pedagogues) in each local office. Each of these supervisors would be in charge of 24 FAMI mothers—a similar caseload as the one we implemented and tested. Similarly, one person in each of the 33 regional offices could by trained by the national office to become trainer of trainers/supervisors.

In our pilot, one senior member of the research team trained a young psychologist in the curriculum, our fieldwork manager). Both trained the 9 tutors during three weeks full-time. After that, the senior researcher was sporadically engaged, but mostly to supervise and provide feedback to the field manager. Each tutor trained approximately 20 to 23 FAMI mothers in sessions that were also 3 weeks long, approximately.

We think it would be feasible to implement a similar cascaded training and supervision process in a potential scale up. Remote online training could be used as a more efficient way of providing pre- and in-service training at the first level of the cascade. Practices by pairs could be supervised online and practices with children could be videotaped for the trainer to provide feedback on the recordings. If required, part of the training could be in-person.

Members of the research team have experience with similar at-scale training strategies for other purposes. In particular, in 2017 we piloted a training scheme that combined presential and online activities for staff in local offices to routinely collect data on child development (the TVIP test) nationwide. Even if connectivity was particularly problematic at the time, this constraint might be easing as digital services expand in response to the COVID- 19 pandemic.

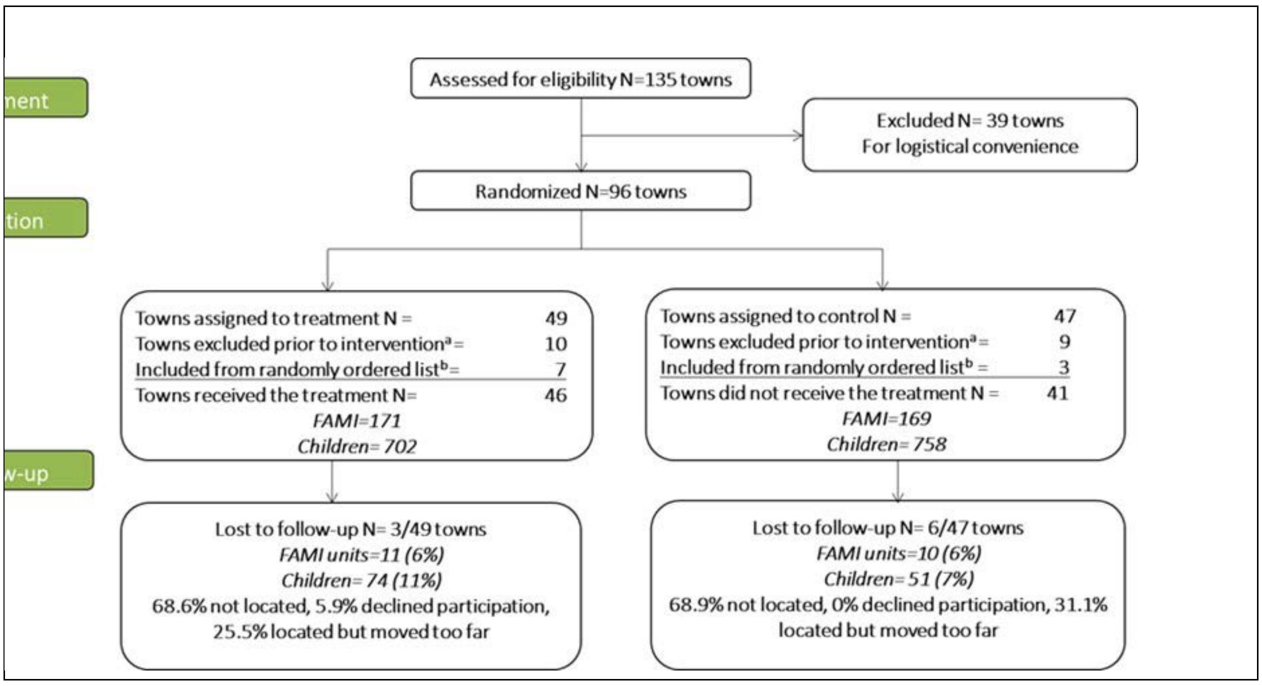
Appendix 3. Study design: power calculations and final sample

Power calculations assumed program effects of 0.25 of a SD relative to the control on the Bayley-III. These were obtained using an average of 4 FAMI units per town and 4 children aged 0-12 months of age per FAMI. We assumed an intra-class correlation within towns of 0.04 (in the Bayley-III scale and conditional on observables), based on the data in Attanasio et al. (2014) for a similar study in Colombia. This sample design provided 95% power at the 5% significance level, allowing for an attrition rate of 10%.

Towns in the final sample had an average of four FAMI units (SD of 2.3, range between 1 and 13), which translates in a total of 171 FAMI units that received treatment and 169 FAMI units that remained as control.

Figure 3.1 presents the c-RCT flow chart, which shows how the final study towns were selected (Figure 3.1). Final 3.2 depicts the final geographic location of the sample.

Figure 3.1. Study's flow chart

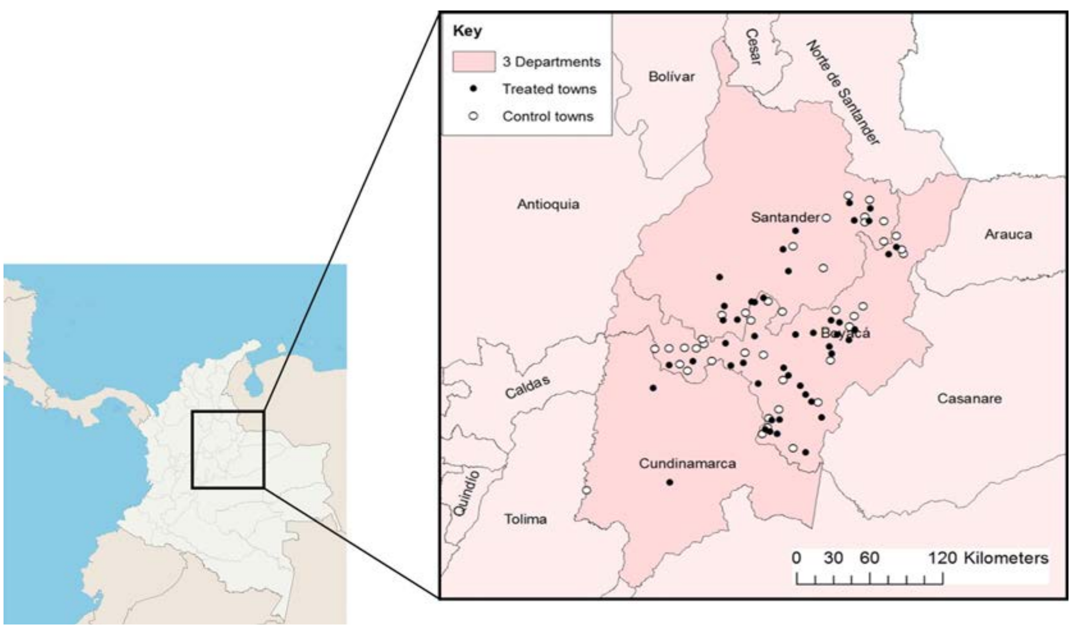


Source: Consort Flow Chart. Own Elaboration

[a] Once in the field for data collection, we realized some towns did not have any FAMI units as they had made the transition to other public parenting programs (*Modalidad Familiar* or MF)

[b] Towns in the list of 39 towns excluded initially from the sample, were randomly ranked and used as replacements. However, we did not have enough replacement towns in all randomization strata.

Figure 3.2. Geographic location of the sample



Note: Treated towns depicted in black and control towns depicted in white.

Appendix 4. Construction of Bayley III-factor and parental investment

As mentioned in Section 3.1, we used Bayley-III age-standardized cognitive, receptive and expressive language, and fine and gross motor to measure child's development. For parental investment, we used the number of magazines, books, or newspapers in the home, the number of toy sources, the number of varieties of play materials in the home, and the number of varieties of play activities the child engaged in with an adult over the three days before the interview. As these are noisy measures of child development and parental investment we follow Cunha and Heckman (2008) and Heckman et al. (2013) and implement a dedicated measurement system that links each observed measure to one latent factor.

We define M_k as the number of measures of the k-th latent factor (e.g., child development or parental investment), and m_{jk} as the j-th measure of the k-th latent factor. Assuming each measure is additively separable in the logarithm of the latent factor we specify:

$$m_{jk} - \mu_{jk} = \alpha_{jk} ln\theta_k + \varepsilon_{jk} \tag{4}$$

where the terms μ_{jk} are the intercept scaled to zero, α_{jk} are the loadings, θ_k the k-th latent factor, and ε_{jk} are the mean zero measurement error terms assumed to be independent of the latent factors and from each other. This specification assumes that the measurement system is invariant across treatment status.

As the latent factors are unobserved, they have no natural scale and identification requires normalizations (Anderson, and Rubin, 1956). Then, we set the scale by setting $Var(ln\theta_k) = 1$. Also, for the child's development factor we set the loading of the Bayley-III cognitive scale to one, that is $\alpha_{1CD} = 1$. Finally, we set the location by setting $E[ln\theta_k] = 0$ for $k = \{CD, PI\}$, where CD and PI, refer to child development and parental investment, respectively.

In Table 4.1 we look at the fraction of the variance in each measure that is explained by the variance in signal. All measures are far from having 100 percent of their variance accounted for by signal, which illustrates the usefulness of the latent factor approach.

Table 4.1. Signal-to-noise ratios

Measurement	Signal
Bayley-III factor	
Bayley: Cognitive	0.666
Bayley: Receptive language	0.772
Bayley: Expressive language	0.635
Bayley: Fine motor	0.589
Bayley: Gross motor	0.494
Parental Investment	
FCI: Number of toy sources	0.161
FCI: Number of types of play materials	0.697
FCI: Number of types of play activities in last 3 days	0.704
FCI: Number of books, magazines, or newspapers	0.332

Note: This table shows the fraction of the variance in each measure that is explained by the variance in signal.

Appendix 5. Descriptive statistics for FAMI facilitators

Table 5.1. Baseline characteristics of FAMI program facilitators by randomization status

	Treatment		Control			
Variables	Mean	SD	Mean	SD	p-value	RW
Age	41.80	(10.04)	41.40	(10.36)	0.790	0.998
Education (years)	13.30	(1.66)	13.00	(1.96)	0.379	0.992
Work experience (years)	11.70	(7.96)	11.90	(8.48)	0.856	0.998
Number of children	2.70	(1.35)	2.50	(1.50)	0.308	0.986
MC's household size	3.90	(1.48)	3.90	(1.43)	0.950	0.998
Number of children (0-12 months) attending	4.80	(2.06)	5.10	(2.29)	0.505	0.996
Number of pregnant women attending	1.80	(1.34)	1.90	(1.45)	0.588	0.997
Number of group sessions held last month	5.40	(4.50)	5.10	(3.39)	0.608	0.997
Number of home visits held last month	12.10	(6.66)	13.50	(7.12)	0.210	0.997
Hours devoted to planning activities (hours)	4.90	(3.02)	6.80	(6.92)	0.014**	0.195
Peabody Picture Vocabulary Test (Z-score)	0.16	(1.03)	-0.17	(0.94)	0.062*	0.610
Knowledge about ECD (Raw Score: correct)	7.29	(1.72)	7.11	(1.39)	0.384	0.992
Single, divorced, or widowed (%)	24	-	21	-	0.555	0.997
No. of observations	171		169			

Note: ***p<0.01; **p<0.05; *p<0.1. Standard deviations clustered by town in parenthesis.
Adjusted p-values using the Romano-Wolf (Romano, and Wolf, 2005; Romano, and Wolf, 2016) procedure (2,000 iterations, clustered by town) are included in the last column. All variables in the table are considered as one group of hypotheses.

Appendix 6. Attrition, compliance, and dosage analysis

Table 6.1. Attrition analysis

Explanatory variables	**Dependent variable -> Lost at FU**		
ITT	0.0383*	0.0401*	-0.0283
	(0.0213)	(0.0208)	(0.0462)
Age at BL (in months)		0.0159*	0.0139
		(0.0088)	(0.0092)
Age squared		-0.001	-0.0012
		(0.0008)	(0.0008)
Child Gender		-0.0176	
		(0.0127)	
First Born		0.0394**	0.0277
		(0.0165)	(0.0253)
High Household Wealth		-0.0279*	-0.0028
		(0.0148)	(0.0191)
Maternal Years of Education		-0.0011	
		(0.0026)	
Father is Present		-0.0450**	-0.0585**
		(0.0208)	(0.0269)
Household Size		-0.0094	
		(0.0066)	
Maternal PPVT		-0.0018	
		(0.0011)	
ITT * Age			0.0090**
			(0.0039)
ITT * First Born			0.04
			(0.0337)
ITT * High Household Wealth			-0.0524*
			(0.0284)
ITT * Father is Present			0.029
			(0.0411)
Constant	0.0675***	0.1408***	0.0718**
	(0.0116)	(0.0469)	(0.0344)
Observations	1,456	1,456	1,456
R-squared	0.0047	0.0334	0.0334
F-stat	3.229	2.908	3.281
Prob > F	0.076	0.004	0.001

Note: *** p<0.01, ** p<0.05, * p<0.1; Standard errors clustered by town in parenthesis.

Attrition. In Table 6.1, we analyze attrition between baseline and follow up. The table indicates that children in treated FAMIs were more likely to be lost at follow up, although the size of this impact is only marginally significant, both with and without controls for a number of baseline characteristics.

Compliance. In Table 6.2 we look at the possibility that the intervention affected the probability of leaving FAMI for alternative programs. We find a significant reduction of leaving FAMI. This result is robust across different specifications (with and without controls) and to different statistical models. In the table we report estimates obtained with a Linear Probability Model, a Logit and a Probit model.

Table 6.2. Compliance analysis: impact of the program on the probability of leaving FAMI for an alternative ECD program

Explanatory Variables	Dependent variable-> Change from FAMI to other ECD program between BL and FU					
ITT	**-0.100***	**-0.125****	**-0.100***	**-0.121****	**-0.100***	**-0.126****
	(0,056)	**(0,051)**	**(0,056)**	**(0,051)**	**(0,056)**	**(0,051)**
Child Gender		0,007		0,009		0,009
		(0,028)		(0,027)		(0,027)
Attendance to ECD at BL		0,032		0,037		0,041
		(0,056)		(0,058)		(0,059)
Town's population range		-0.104*		-0.103*		-0.105*
		(0,055)		(0,054)		(0,056)
High Household Wealth		0.070**		0.072**		0.073**
		(0,034)		(0,034)		(0,033)
Teenage mother		-0,015		-0,013		-0,017
		(0,033)		(0,033)		(0,033)
Maternal PPVT		0.012***		0.011***		0.011***
		(0,002)		(0,002)		(0,002)
Department FE		-0,022		-0,022		-0,017
		(0,061)		(0,064)		(0,064)
Observations	989	989	989	989	989	989
Model	LPM	LPM	Probit	Probit	Logit	Logit

Note: *** p<0.01, ** p<0.05, * p<0.1; Standard errors clustered by town in parenthesis. Average marginal effects presented.

In Table 6.3 we look at the effect of the intervention on the probability of leaving FAMI to join *Modalidad Familiar*. We do not find any evidence of a significant effect.

Table 6.3. Compliance analysis: impact of the program on the probability of leaving FAMI for *Modalidad Familiar* (MF)

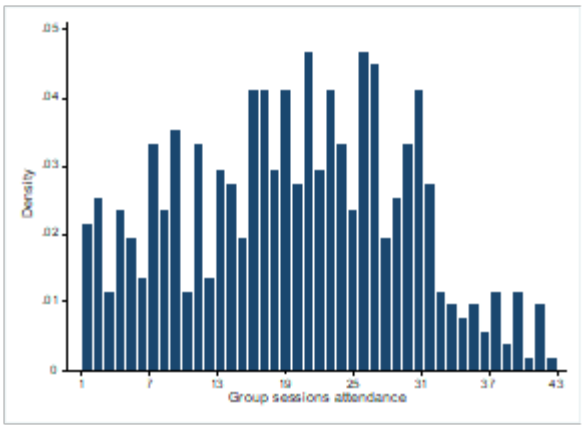
Explanatory variables	ent variable-> Change from FAMI to MF between BL and FU					
ITT	**-0,029**	**-0,058**	**-0,029**	**-0,06**	**-0,029**	**-0,06**
	(0,048)	**(0,045)**	**(0,048)**	**(0,046)**	**(0,048)**	**(0,047)**
Child Gender		-0,024		-0,021		-0,022
		(0,021)		(0,020)		(0,021)
Attendance to ECD at BL		0,022		0,034		0,032
		(0,042)		(0,047)		(0,049)
Town's population range		-0,038		-0,043		-0,041
		(0,049)		(0,049)		(0,053)
High Household Wealth		0,024		0,026		0,025
		(0,025)		(0,024)		(0,024)
Teenage mother		-0,031		-0,031		-0,032
		(0,026)		(0,025)		(0,025)
Maternal PPVT		0.008***		0.007***		0.007***
		(0,003)		(0,002)		(0,002)
Department FE		0,018		0,026		0,025
		(0,039)		(0,042)		(0,041)
Observations	770	770	770	770	770	770
Model	LPM	LPM	Probit	Probit	Logit	Logit

Note: *** p<0.01, ** p<0.05, * p<0.1; Standard errors clustered by town in parenthesis. Average marginal effects presented.

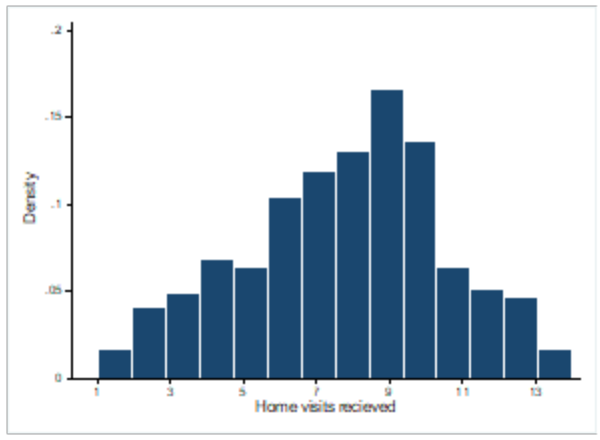
Dosage. The following set of Figures show the distribution of children in the intervention group by the duration of exposure to the pedagogical component of the program; as well as joint compliance between the nutritional and pedagogical components. Table 6.4 compares household characteristics by total exposure to the program.

Figure 6.1. Effective individual program participation

a. Attendance to group sessions

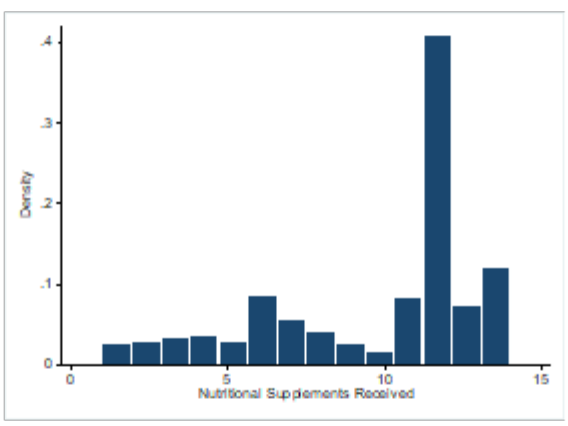


b. Attendance to group sessions

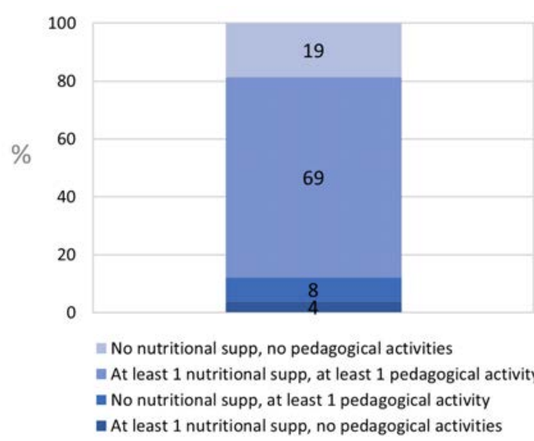


Source: Program attendance registry (recorded by FAMI facilitators)
Notes: a. Subsample of children registered at least once in group session attendance lists (74.2% of all treated children assessed at baseline). b. Subsample of children registered at least once in home visit attendance lists (71.7% of all treated children assessed at baseline).

c. Nutritional supplement



d. Joint compliance with the nutritional and pedagogical components



Source: Fundación Exito records.
Maximum nutritional supplements during the intervention period=14
Pedagogical activity: group session or home visit.

Table 6.4 Differences in child and household characteristics by duration of exposure to the program

	Attendance above median		*Attendance below median*		
	Mean	SD	Mean	SD	
Sociodemographic characteristics					
Child's age in months	5.43	(3.31)	6.11	(3.55)	***
Boy (%)	0.53	(0.50)	0.50	(0.50)	
Child's birthweight (gr)	3197	(585)	3170	(509)	
Birth order	1.06	(1.34)	1.01	(1.22)	
First born (%)	0.43	(0.50)	0.50	(0.50)	
Maternal years of schooling	8.58	(3.48)	9.14	(3.34)	*
Father present (%)	0.74	(0.44)	0.66	(0.48)	**
Number of siblings	1.03	(1.26)	1.02	(1.36)	
Mother married / cohabiting	0.71	(0.45)	0.71	(0.45)	
Mother is single	0.22	(0.41)	0.27	(0.45)	
Mother is divorced	0.01	(0.12)	0.01	(0.09)	
Teenage mothers (%)	0.23	(0.42)	0.28	(0.45)	
Mother's age (years)	26.79	(7.00)	25.48	(6.58)	**
Household wealth index [a]	0.02	(0.96)	0.10	(0.96)	
No. of observations	366		336		
Intermediate Outcomes					
FCI No. of adult books, magazines and newspapers	2.54	(3.07)	2.60	(3.10)	
FCI No. of toy sources	1.33	(0.94)	1.38	(0.93)	
FCI No. of varieties of play materials	1.33	(1.37)	1.52	(1.44)	
FCI No. of varieties of play activities over past 3 days	2.45	(1.51)	2.62	(1.60)	
FCI No. of parental care activities over past 3 days	4.65	(1.13)	4.86	(0.91)	***
Parental Investment [b]	-0.08	(0.98)	0.07	(0.98)	*
Social support DUKE UNC-11 total (raw score)	40.75	(8.39)	41.74	(8.10)	
High maternal self-efficacy (% above median)	0.39	(0.49)	0.41	(0.49)	
Mothers with depression symptoms (%)	0.14	(0.35)	0.17	(0.37)	
Use of verbal or physical abuse in the household (%)	0.01	(0.10)	0.04	(0.19)	*
No. of observations	366		336		
Nutritional Status					
Weight-for-age z-score (n1=351, n0=311)	0,18	(1,42)	0,33	(1,36)	
Length/height-for-age z-score (n1=345, n0=309)	0,00	(1,66)	-0,04	(1,74)	
BMI-for-age z-score (n1=338, n0=299)	0,33	(1,64)	0,39	(1,62)	
Weight-for-length/height z-score (n1=331, n0=295)	0,41	(1,56)	0,32	(1,62)	
Food-Insecure (%)	0,47	(0,50)	0,54	(0,50)	

Note: ***p<0.01; **p<0.05; *p<0.1. Median= 21 activities (groups sessions and home visits).
Standard errors clustered by town in parenthesis
a The wealth index was computed as the principal component of a set of dichotomous variables that describe characteristics of the household, ownership of durable goods, and access to public utilities.
b Factor score of FCI items.

Appendix 7. Main impacts (ITT) corrected for attrition

In Table 7.1, we report the estimates of the impact we obtain by modeling attrition as a selection process. The identifying assumption of the impact results is that the quality and motivation of the interviewers (randomly assigned to both treatment and control towns) determines the probability of attrition. We find that interviewers dummies are a significant determinant of attrition and that the results on the impacts of the program are virtually unaffected by non-random attrition.

Table 7.1 Program impacts on children's outcomes estimated by maximum likelihood correcting for self-selection into the follow up sample

VARIABLE	Impacts (95% CI)	P Value	RW P Value
Bayley-III Factor	0.159 (0.030,0.289)	0.009***	0.027**
ASQ:SE Total Score	-0.075 (-0.250,0.099)	0.398	0.579
Height-for-age Z-Score	0.075 (-0.075,0.187)	0.194	0.400

Note: ***p<0.01; **p<0.05; *p<0.1; 95%. Confidence interval in parenthesis for two-tailed tests. Standard errors clustered by town. p-values are computed using Romano-Wolf (2005, 2016) step-down procedure. We consider 3 hypotheses for children's outcomes. Exclusion restrictions: interviewer fixed effects at baseline and assigned interviewer fixed effects at follow-up. First stage F-stat=11.24. Bayley-III factor is the factor score of the age-standardized Bayley III scales. ASQ:SE Total Score is the age-standardized ASQ:SE score.

Appendix 8. Evaluation of an Integrated Intervention Targeted at Deprived Pre-School Children in Rural Colombia Pre-Analysis Plan. 22/06/2016

1. Introduction

This document outlines a pre-analysis plan—study design, hypothesis to be tested, and data and specifications to be used—for evaluating the impact of the integration of a structured curriculum (both, home visits and group meetings) to promote young children's development into the FAMI (*Hogares Comunitarios - Modalidad Familia, Mujer e Infancia)* parenting programme for disadvantaged families in rural Colombia. The program is being implemented by the research team in cooperation with the National Family Welfare Agency (*Instituto Colombiano de Bienestar Familiar* - ICBF). The intervention and evaluation have been funded by Grand Challenges Canada (GCC) and the Fundación Éxito (FE).

Baseline data collection took place between August and November 2014. Follow-up data will take place between April and July 2016. The intervention ran from September 2014 through March 2016. This plan has been written up prior to follow-up data processing, serving as a pre-commitment for subsequent analysis.

This document is structured as follows. Section 2 reviews the intervention and evaluation design. Section 3 enumerates the hypotheses to be tested as part of the study and the data we will use to test them. Finally, Section 4 outlines the empirical specification(s) to be used in analyzing the data and other data management issues.

2. Overview of the Study: Interventions and Evaluation Design

Hogares Comunitarios – Modalidad Familia, Mujer e Infancia (FAMIs): are small-sized community centers located in areas of high social and economic vulnerability in semi-urban and rural areas of Colombia where pregnant women and parents of children younger than two years of age receive training regarding parental practices including family relationships, pregnancy, breastfeeding, nutrition, health and the upbringing of young children. The program targets socioeconomically vulnerable pregnant women, nursing mothers and parents of children less than 2 years of age. Program eligibility is defined by the national proxy means test (PMT) known as SISBEN, which classifies households into socioeconomic vulnerability levels based on a household survey. A front line worker known as *Community Mother* (MC) works 80 hours per month for the program. In particular, she devotes 32 hours for parental training in group

sessions; 20 hours of home visits (minimum 1 visit per family per month); 8 hours of training for the MC; and 20 hours for planning activities, documentation and transportation times. Group sessions are held separately by age subgroups according to the child's age: children from 0 to 5 months old, children from 6 to 11 months old, 1 to 2 years old children and pregnant women. Each FAMI unit has an average of 12 to 15 beneficiaries.
The FAMI program has two main components: i) the provision of nutritional supplement – that should cover 20 to 25% of daily nutritional requirements of the child or the pregnant women; and ii) training of beneficiary families on parental practices and child development since gestation and up to age two, particularly regarding nutrition, socioemotional development, health, maternal health, and early stimulation through group sessions and one-on-one home visits.

Our study offers a rigorous evaluation, by Randomized Controlled Trial (RCT), of the short-term impacts of <u>the integration of a structured curriculum (both, home visits and group meetings) to the FAMI program and the addition of rigorous training and supervision protocols for front line workers</u>.

The Intervention
Based on a rigorous study on the weaknesses of the program and a small pilot on the improvements that were to be developed, the team designed an **integrated upgrade** of the FAMI program. The upgrade consists of 3 components. First, the implementation of a structured but *feasible*, *flexible* and *effective* curriculum (both for home visits and group meetings) focused on encouraging mother-child interactions and maternal self-efficacy, teaching mothers how to promote their children's development and promoting maternal self-esteem and mental health. This component is complemented by the provision of pedagogical materials such as puzzles and books, materials for home-made toys and toy making workshops during group sessions. Second, the improvement of MCs' training, provided by professional tutors trained by the research team, in order to guarantee the fidelity of program implementation and the addition of a supervision and coaching protocol for front line workers to be delivered throughout the duration of the intervention. Third, the delivery of an additional nutritional supplement (increased intake of calorie, protein, vitamins and minerals), complemented with psychoeducation around feeding and nutrition during sessions, and informative materials to promote healthy nutritional habits.

Sample and Evaluation Design
The evaluation sample consists of 1,466 children 0-12 months at baseline and 553 pregnant women in 340 FAMIs located in 87 municipalities of three Colombian departments of the Andean region: Boyacá, Santander and Cundinamarca. These municipalities were selected for their geographical location, their semi-rurality and rurality conditions and the presence (or not) of other similar ECD services such as another parenting program similar to the one included in this study, known as MF or *modalidad familiar*.
From a universe of 151 eligible municipalities (i.e. municipalities with less than 40.000 inhabitants, at most one MF and that belong to Boyacá, Santander, Cundinamarca and Tolima departments), we selected all municipalities with no MFs and then we selected from the remaining municipalities with at most one MF, striving to achieve distributed geographic coverage, until we reached 96 town. Then we ran a stratified randomization based on (i) MF (*modalidad familiar*) presence, (ii) Department and (iii) Population size (less than or more than 10.000 inhabitants) resulting in 49 municipalities in the treatment group and 44 in the control group. We then dropped FAMIs that were transiting or were going to transit to the new version of the program (MF) and, as a consequence, had to drop all study municipalities located in the department of Tolima since no control municipalities were left for this group. This procedure implies a final sample of 41 municipalities with 169 FAMIs in the control group and 46 municipalities with 171 FAMIs in the treatment group.
Baseline data collection on the children and the pregnant women, their households, the MCs and the centers they attend took place between August and November 2014. Follow-up data collection will take place from April to July 2016. We collected baseline data directly in participants' homes, except in those instances in which it was not possible to interview the mother of the child in her own home, in which case the interview took place in community centers at the town's urban center such as schools, churches or in the FAMI.

The analysis of baseline data shows that the sample is balanced across the evaluation groups. Whilst there are some significant differences in children's nutritional status and a few socio-demographic characteristics (e.g. presence of the father and mother's TVIP and personality), and in some of sociodemographic characteristics of MCs, none of these differences systematically occur in one of the study groups nor point towards a specific (bias) direction.

3. Hypotheses to be Tested and Data

We have collected (at baseline) and are collecting (at follow-up) a rich set of data and we will use it to test a series of hypotheses concerning the impacts of the intervention under study. We present the study hypotheses in two groups: impact on children's outcomes and impact on mothers' parenting abilities and on the learning environment at home.

I. Hypotheses Group A: Impact on Children's Outcomes

The treatment may have positive average impacts on outcomes for children attending the program. We group these outcomes into two areas: children's development and children's nutritional status.

We consider a number of domains within development—namely, some aspects of cognition, language and motor and socio-emotional development. We next list the specific hypothesis on each of them by domain of development, and detail the specific tests (and scales) we will use to measure them and how we will process the data. We would however like to flag two considerations before proceeding.

On the one hand, we would like to clarify that we plan to use **factor analysis** (on standardized scores) to determine the most appropriate way of combining the various tests and scales collected in "constructs". The reason for this is that child development is composed of many different dimensions that are interrelated. For example, even if a vocabulary test should be viewed as an achievement test as opposed to a measure of raw ability (given that it measures acquired vocabulary) it nevertheless correlates well with ability. Hence, it is very difficult to establish a priori the most sensible way to organize the data, and we plan to rely on factor analysis to combine data that captures common underlying constructs (data that would be thought to go together on theoretical grounds). To do this, we will follow standard protocols in the use of factor analysis. First, we will construct as many factors ("constructs") as there are with eigenvalues larger than 1. Next, we will only use outcomes (scales or tests) with factor loads larger than 0.4 in the construction of these factors. Hence, the following categorization of scales and tests in domains is for the purpose of illustrating our hypothesis and may be modified as a result of the outcomes of factor analysis.

In addition to the main analysis based on the impacts on the factors we have identified, we will also report impacts on the individual tests, correcting our p-values for multiple hypotheses testing, using the Romano-Wolf step-down procedure (Romano and Wolf 2005a,b), as further discussed in Section 5.

On the other hand, note that the following child development outcomes, listed under Hypotheses Group A, will be collected at follow-up only (as they were not suitable for administration to children at baseline given their age) by direct administration to children by a trained psychologist at a community venue in each town. An expert psychologist with extensive training and practice in the collection of the assessment instruments included in this study trained a group of 10 psychologists who administered the assessments in the field. The only exception to this is the ASQ:SE, which is collected by maternal report (direct interview with the mother, and was included as part of the household questionnaire). Children were 0-12 months of age at baseline and will be 15-29 months of age at follow up, depending on the exact time at which they are assessed at follow-up. At baseline we collected different child outcomes, which will be used as baseline controls, as explained in Section 4.

Hypothesis A1: *The treatment is likely to have a positive average impact on children's cognitive development, language development, and motor development.*

These domains will be assessed using the indicators listed next.

- Cognitive development: some dimensions of cognition will be assessed by the Spanish Version of the *Cognitive Scale* of the **Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)** (Bayley, 2006). The scale evaluates the sensor and motor development, exploration and manipulation, the relationship of objects, concept formation, memory, and other aspects of cognitive processing of the child.

- o The scales will be administered and scored as indicated in the Bayley-III administration manual. Higher scores indicate higher cognitive abilities. However, given the process of development, scores are also likely to increase with age.

- Language development: we will assess both receptive and expressive language.
- Receptive language will be assessed using the Spanish version of the *Receptive Communication Scale* of the **Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)** (Bayley, 2006). The scale evaluates pre-verbal behaviors and vocabulary development – i.e. the ability to identify objects and images that are been referenced.
- o The scale will be administered and scored as indicated in the Bayley-III administration manual. Higher scores indicate higher receptive language development. However, given the process of development, scores are also likely to increase with age.

- Expressive language will be assessed using the Spanish version of the *Expressive Communication Scale* of the **Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)** (Bayley, 2006)The scale evaluates pre-verbal communication, such as babbling, gesticulation, joint referencing and early talking; and vocabulary development, such as naming objects, images and attributes (eg, color and size).
- o The scale will be administered and scored as indicated in the Bayley-III administration manual. Higher scores indicate higher expressive language development. However, given the process of development, scores are also likely to increase with age.

- Motor Development: we will assess both fine and gross motor development.
- Fine Motor Development will be assessed using the Spanish version of the *Fine Motor subset Scale* of the **Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)** (Bayley, 2006). The scale evaluates the child's skills related to visual tracking, reaching, gripping and manipulation of objects. Functional abilities of the child's hand and responses to tactile information are also measured.
- o The scale will be administered and scored as indicated in the Bayley-III administration manual. Higher scores indicate higher fine motor development. However, given the process of development, scores are also likely to increase with age.
- Gross Motor Development will be assessed using the Spanish version of the *Gross Motor subset Scale* of the **Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III)** (Bayley, 2006). The scale evaluates the child's static position (e.g. when the child is sitting or standing); dynamic movement, including locomotion and coordination; balance; and planning of movements.
- o The scale will be administered and scored as indicated in the Bayley-III administration manual. Higher scores indicate higher gross motor development. However, given the process of development, scores are also likely to increase with age.

Hypothesis A2: *The treatment is likely to have a positive average impact on children's socio-emotional development.*

- Socio-emotional development: we will assess socio-emotional development using the **Ages and Stages: Socio-Emotional Questionnaire** (ASQ:SE), which screens several socio-emotional areas such as self-regulation, compliance, communication, adaptive behaviors, autonomy, affect, and interaction with people, for children 15-29 months at follow-up by parental report.
- o The scale will be administered and scored as indicated in the ASQ:SE manual.
- o While scores should not be age dependent, we will remove any lingering age effect standardizing the scores internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods.

Hypothesis A3: *The treatment is likely to have a positive average impact on children's nutritional status.*

We expect the intervention to improve nutritional status through the addition of a nutritional supplement, the education provided about feeding and nutrition during sessions and home visits, and the provision of informative materials to promote healthy nutritional habits. Children's nutritional status is measured both at baseline and at follow-up by personnel that is trained by an expert nutritionist and assessed for reliability. In particular, we collected information on height, weight and body mass index (BMI) following World Health Organization (WHO) standards (WHO Multicentre Growth Reference Study Group, 2006, 2007) for all children in our sample, both at baseline and follow-up. Based on these measures, we will construct a variety of nutritional indicators depending on the child's age and based on WHO (2006, 2007) standards.

Hypothesis A4: *The treatment is likely to have a positive average impact on children's food insecurity status.*

We expect the intervention to improve the household's and the child's food insecurity status as a result of the nutritional supplement delivered with the program. We assess food insecurity status using the Latin-American and Caribbean Nutritional and Food Insecurity Scale (ELCSA) adapted and validated for this population.

II. Hypotheses Group B: impact on the mother's parenting skills and the learning environment at home

The intervention is more likely to improve children's outcomes if the children's mothers effectively internalized the message and training delivered by the program's new curriculum and took these lessons to practice with their own children. In particular, the program might have been more effective if maternal self-esteem, mental health and motivation improve as a result of program participation; if the quality of the home environment improves; if discipline practices at home are more appropriate; and if parents devote more time to stimulating activities with their children such as reading and playing. Also, these things are more likely to have happened if parents participated more regularly in the program's group sessions and home visits.

***Common Hypothesis B*:** *We will test the hypotheses that the intervention will have an impact on mother's parenting skills, parental knowledge and perceptions, parental self-efficacy, mental health, and the home environment. We also hypothesize that these changes in parental practices, knowledge, perceptions, self-esteem, mental health, motivation and changes in the learning environment at home will correlate with (and contribute to—i.e. mediate) the impacts on children's outcomes described above (Hypotheses Group A).*

We next describe the outcomes previously listed and how we will construct them.

i. Maternal self-efficacy

To measure maternal efficacy, we use an adapted version of the 15-item Self-Efficacy in the Nurturing Role Questionnaire (Pedersen et al., 1989; Gibaud-Wallston y Wandersman, 1978). Mothers responded *Yes, No* or *Unsure* to indicate how they felt about their role as a parent.

ii. The quality of the home environment is improved

We will use an adaptation of a subset of questions from the UNICEF scale of the home environment quality: The Family Care Indicator (FCI) - Version 16 (Frongillo, E., Sywulka, S., & Kariger, P., 2003) to measure the quality of the home environment. The FCI measures the availability and variety of play materials, sources of these materials and books and magazines in the child's household.

iii. Discipline strategies used by parents

We will use an adaptation of the domain III of the UNICEF Care Indicator Questions (version 16). The questionnaire inquires about the discipline strategies most frequently used by parents at home with their children. In particular, the questions aim at determining how frequent nonviolent strategies, psychological aggression strategies, and physical punishment strategies are used with children.

iv. *Parental time devoted to stimulating activities with their children*

We will use an adaptation of a subset of questions of the Family Care Indicator (FCI) - Version 16 (Frongillo, E., Sywulka, S., & Kariger, P., 2003) which measures the amount of time devoted by adults in the household to a variety of activities, including personal care routines, with the child during the three days prior to the interview date.

v. *Parental knowledge about early childhood development*

We use a 10-item scale to assess maternal factual knowledge of childcare/parenting practices, child developmental processes, and milestones. This scale has been piloted by the research team but has not been formally assessed for reliability and validity.

vi. *Parental mental health*

We use the CESD-10 scale to assess maternal mental health. The 10-item Center for the Epidemiological Studies of Depression Short Form (CES-D-10) is a widely used measure to screen for depression. The 10-item measure has demonstrated strong psychometric properties, including predictive accuracy and high correlations with the original 20-item version. It is generally used to assess depression symptom severity rather than as a diagnostic screening tool (Anderson et al., 1994).

vii. *Parental social support*

We use the DUKE-UNC Functional Social Support Questionnaire 11 to assess the mother's perception about the amount and type of functional social support that she receives. Given the characteristics of the intervention, we seek to understand whether parenting group meetings contribute to a perception of enhanced social support, which might contribute towards maternal self-efficacy and maternal mental health.

viii. *Mothers attend FAMI group sessions and receive home visits.*

We will use self-reported data on attendance and turnover, collected from the main caregiver, and from administrative data on attendance.

4. Use of the qualitative and process data

In addition, as part of a qualitative evaluation, we are collecting process data on a subsample of FAMIs that will contribute to inform the quality (and fidelity) of the program, and provide insights on the extent to which the mechanisms through which the program has (or did not have) impacts occurred. The objective of this qualitative component is to characterize the interventions in the field with a detailed tracking of activities and responsible individuals related to the components of the interventions under study. Based on direct observations, in-depth interviews and session videotaping we hope to gain a better understanding of the ways in which the interventions are understood, adopted and used directly by MCs. In addition, interviews to beneficiary mothers are taking place, to capture mother's perceptions about the program's strengths and weakness.
With this input we hope to be able to (1) better interpret our quantitative estimates of program impacts, (2) identify possible transmission mechanisms of the effects of interest, and (3) propose specific policy recommendations for program improvement based on the joint analysis of the quantitative and the qualitative results.

5. Empirical Strategy

Baseline Specification

Given the experimental design described in Section 2, we can identify the impacts of the treatment on outcomes using the following estimating equation:

$$Y_{isl,1} = \beta_0 + \beta_1 T_{sl} + \gamma Y_{isl,o} + X'_{isl,0}\delta + D_{isl,0}\theta + F_{isl,0}\sigma + S_{isl,0}\tau + Z_{isl,1}\rho + \varepsilon_{isl,1} \qquad (1)$$

where$Y_{isl,1}$is the outcome of interest for child i in FAMI center s in municipality l at follow-up (t=1); T_{1sl} is a dummy equal to 1 if the FAMI center s in municipality l receives the treatment; and $Y_{isl,0}$ is the baseline (t=0) level of the outcome of interest (or level of the corresponding aggregate construct in the case that the same measure was not administered at baseline and follow-up) for child i in FAMI center s in municipality l at follow-up. For child developmental outcomes we will not have the same outcome at baseline and follow-up since the tests could not be administered given children's ages at baseline. For these outcomes, we will control for all existing aggregate scores (constructed using factor analysis, as described above) and including all developmental scores and nutritional scores. The purpose of this approach is to maximize efficiency. $X'_{isl,0}$ is a set of basic child and household characteristics, which are also added to improve efficiency (minimize residual variance) and control for the slight imbalance in some baseline characteristics observed between groups at baseline (detailed list pending); $D_{isl,0}$ are a set of department fixed effects, $F_{isl,0}$ are a set of dummies indicating the presence or not of the alternative parenting program in town l (*modalidad familiar)* and $S_{isl,0}$ are a set of municipality population size dummy variables indicating above and below 10 thousand inhabitants (all included due to our stratified randomization procedure), and $Z_{isl,1}$ are a complete set of tester or interviewer dummies. $\varepsilon_{isl,1}$ is the random error term, clustered at the municipal level l (the unit of randomization). 32

We can estimate equation (1) by OLS. β_1 is the estimated average impact of the treatment on outcome $Y_{isl,1}$; (intent-to-treat estimate). If compliance is not complete in the sense that children do not attend all program sessions they are intended to (and assuming this non-compliance is not larger than 40%) we would additionally estimate duration of exposure to treatment effects by instrumenting actual duration of participation in the program with the result of the random assignment (intention-to-treat or randomized treatment variable). We would also assess the extent to which actual duration of exposure to treatment is correlated with treatment status.

Impact on mother's parenting skills and the home environment (Hypotheses Set B):

We can also use equation (1) to estimate the impact of the treatment on intermediate outcomes. When the impact on mother's parenting skills and the home environment refer to the child's mother (or other caregiver) we will replace the set of basic covariates in X with mother (or other caregiver) basic baseline or time invariant characteristics (e.g., age and educational attainment).

Dealing with Testing for Multiple Outcomes through Standardized Treatment Effects and Adjustments for Multiple Inference

For some of the developmental domains analyzed in this study, we have more than one outcome measure with which to explore treatment effects. To deal with multiple hypothesis testing we will employ two approaches.

The first approach will be to group our outcome measures into domains or "constructs" using factor analysis (following the procedure described in Section 3) and estimate equation (1) using the resulting factor index as the relevant dependent variable. This procedure is based on the idea that items within a domain are measuring an underlying common "construct" (or factor).

The second approach will consist of estimating each outcome (individual test) independently but adjusting p-values for multiple hypotheses testing using the step-down procedure developed by Romano-Wolf on each set of (Romano and Wolf, 2005).

Survey attrition

We acknowledge that a certain level of attrition is unavoidable. We will check that the sample of non-attriters remains balanced on baseline observables (as is the entire sample). We will also check that attrition is independent of treatment status. In the event of a significant correlation

[32] Note that as standard errors are being clustered at the municipality level, it is not necessary to account for additional clustering at the FAMI level. Robust standard errors at the larger cluster are less favorable in terms of evaluating our null hypothesis.

between attrition and treatment status we will estimate the determinants of attrition using a probit regression on observables and if feasible use a Heckman selection correction procedure to adjust the estimates of the main equation (1).

Procedures for Addressing Missing Data

We will not impute the values for any dependent variable (final or intermediate outcomes) at follow-up. Regarding missing data on covariates, $Y_{isl,o}$ and $X'_{isl,o}$, we will check whether item non-response is correlated with treatment status. If it is not correlated, we will impute the missing covariate value with the average of the non-missing observations and this imputation will be accounted for with a dummy variable (we will check the robustness of our results by also estimating the regression without that covariate). If non-response in the baseline covariate is correlated with treatment status, we will not use that covariate when estimating the regressions. In cases in which the percentage of observations with covariate missing data is less than 2%, we will simply work with the sample with non-missing data.

Questions with Limited Variation

We will not use as dependent or independent variables any indicator variable that has a prevalence rate of below 10% or above 90%, in order to limit noise caused by variables with minimal variation.

In the event that omission decisions result in the exclusion of all constituent variables (or for as many as indicated in the test manual) for an indicator, the indicator will not be calculated.

Treatment of Outliers

We will drop children with developmental outcomes or nutritional status with standardized values lower than 3 standard deviations below the mean (<-3SD) of the relevant standardized distribution, since we consider this to be an indication of potential disability (for developmental outcomes), severe malnutrition (for nutritional status) or significant measurement error.

References

Andersen E. M., Malmgren J. A., Carter W. B., Patrick D. L. (1994). Screening for depression in well older adults: Evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). American Journal of Preventive Medicine, 10, 77-84.

Bayley N. (2006) Bayley Scales of infant development: manual, 2nd ed. San Antonio, TX: Psychological Corporation.

Frongillo, E., Sywulka, S., & Kariger, P. (2003). UNICEF psychosocial care indicators project. Final report to UNICEF. Ithaca: Division of Nutritional Sciences, Cornell University.

Gibaud-Wallston A, Wandersman LP. (1978) Development and utility of the Parenting Sense of Competence Scale. Paper presented at the meeting of the American Psychological Association; Toronto, Canada.

MacPhee, D. (1981). Manual: Knowledge of Infant Development Inventory, unpublished manuscript, University of North Carolina.

Pedersen FA, Bryan YE, Huffman L, Del Carmen R. (1989) Constructions of self and offspring in the pregnancy and early infancy periods. Paper presented at the Society for Research in Child Development; Kansas City, MO.

Romano JP, Wolf M. 2005. Stepwise Multiple Testing as Formalized Data Snooping. Econometrica; 73(4):1237-1282.

Romano JP and Wolf M. 2005. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. Journal of the American Statistical Association; 100(469):94-108.

WHO Multicentre Growth Reference Study Group (2006). WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development, p. 312. Geneva: World Health Organization.

WHO Multicentre Growth Reference Study Group (2007). WHO Child Growth Standards: Head circumference-for-age, arm circumference-for-age, triceps skinfold-for-age and subscapular skinfold-for-age: Methods and development. Geneva: World Health Organization.