Alison Andrew
Orazio Attanasio
Raquel Bernal
Lina Cordona
Sonya Krutikova
Diana Martinez Heredia
Carlos Medina
Ximena Peña
Marta Rubio-Codina
Marcos Vera-Hernandez

# Evaluation of infant development centres

## An early years intervention in Colombia

June 2018

Impact Evaluation Report 76

Education

3ie 10 YEARS
International Initiative for Impact Evaluation

## About 3ie

The International Initiative for Impact Evaluation (3ie) is an international grant-making NGO promoting evidence-informed development policies and programmes. We are the global leader in funding, producing and synthesising high-quality evidence of what works, for whom, how, why and at what cost. We believe that using better and policy-relevant evidence helps to make development more effective and improve people's lives.

## 3ie impact evaluations

3ie-supported impact evaluations assess the difference a development intervention has made to social and economic outcomes. 3ie is committed to funding rigorous evaluations that include a theory-based design and use the most appropriate mix of methods to capture outcomes that are useful in complex developmental contexts.

## About this report

3ie accepted the final version of the report, *Evaluation of infant development centres: an early years intervention in Colombia,* as partial fulfilment of requirements under grant OW4.1162 awarded under Open Window 4. The content has been copy-edited and formatted for publication by 3ie.

The 3ie technical quality assurance team for this report comprises Diana Lopez-Alva, Radhika Menon, an anonymous external impact evaluation design expert reviewer and an anonymous external sector expert reviewer, with overall technical supervision by Jyotsna Puri. The 3ie editorial production team for this report comprises Angel Kharya and Akarsh Gupta, with Beryl Leach providing overall editorial supervision.

All the content is the sole responsibility of the authors and does not represent the opinions of 3ie, its donors or its board of commissioners. Any errors and omissions are also the sole responsibility of the authors. All affiliations of the authors listed in the title page are those that were in effect at the time the report was accepted. Any comments or queries should be directed to the corresponding author, Sonya Krutikova, at sonya_k@ifs.org.uk.

3ie has received funding for the Policy Window from UK aid, the Bill & Melinda Gates Foundation and the William and Flora Hewlett Foundation, as well as other institutional donors. A complete listing of all of 3ie's donors can be found on the 3ie website.

Suggested citation: Andrew, A, Attanasio, O, Bernal, R, Cordona, L, Krutikova, S, Heredia, DM, Medina, C, Peña, X, Rubio-Codina, M and Vera-Hernandez, M, 2018. *Evaluation of infant development centres: an early years intervention in Colombia,* 3ie Impact Evaluation Report 76. New Delhi: International Initiative for Impact Evaluation (3ie). Available at: doi: https://doi.org/10.23846/OW4.IE76

Cover photo: EC/ECHO/ Boris Heger

© International Initiative for Impact Evaluation (3ie), 2018

# Evaluation of infant development centres: an early years intervention in Colombia

Alison Andrew
The Institute for Fiscal Studies

Orazio Attanasio
The Institute for Fiscal Studies

Raquel Bernal
Universidad de Los Andes

Lina Cordona
Banco de la República, Medellín, Colombia

Sonya Krutikova
The Institute for Fiscal Studies

Diana Martinez Heredia
Universidad de Los Andes

Carlos Medina
Banco de la República, Colombia

Ximena Peña
Universidad de Los Andes

Marta Rubio-Codina
Inter-American Development Bank

Marcos Vera-Hernandez
University College London

**3ie Impact Evaluation Report 76**

**June 2018**

## Summary

This report evaluates the impact of a programme implemented by the government of Colombia to upgrade pre-schools targeted at children from lower socio-economic groups. These pre-schools, called *Hogares Infantiles* (HIs), are run by the Instituto Colombiano de Bienestar Familiar and there are 1,008 of them across Colombia. The pre-schools are community-based and typically located in fairly well-equipped community centres and run by parent associations. The upgrade of HIs to "*Hogares Infantiles Mejorados*" (HIM) was implemented as part of the '*De Cero a 5iempre*' and entailed a considerable financial commitment from the government. It planned to (a) hire additional and better-qualified personnel (including nutritionists, experts in child socio-emotional development and pedagogical assistants), as well as (b) deliver a one-time pedagogical endowment valued at US$52 per child for toys, books and other materials. In addition to the HIM upgrade, some HI centres received a set of further improvements, provided by a private foundation, *Fundación Éxito* (FE). These improvements (HIM+FE) included three core programmes - further nutritional improvement, pedagogical training of teachers and a reading programme for teachers, children and parents.

In collaboration with the government, we designed a controlled experiment. Our design allowed us to evaluate the effects of the government-implemented improvements alone (HIM), as well as of the combined programme, which includes the add-on effects of the enhancements delivered by FE. The study took place in the eight largest cities in Colombia.

From the 670 HIs in the 8 cities, we selected 120 to include in the study. These were organised into 40 groups of 3 by geographic proximity. Within each group, we randomly assigned one HI to each of the control group, the HIM group and the HIM+FE group. We randomly selected 17 children between 12 and 36 months at baseline from each HI. We collected data on the 120 HIs and 1,989 children and their households, both before the enhancement of HIs (baseline, March-June 2013) and 18 months after (follow-up, October-November 2014).

Implementation data showed that not all HIMs fully implemented the programme. For the HIM intervention, just 34 of the 80 HIM centres hired at least half of the planned number of professionals (on a full time equivalent basis) for all three categories of staff. For the FE intervention, there was variation in the proportion of teachers who attended the FE pedagogical training with a third or more teaching staff participating in 25 of the 40 centres.

The main outcomes of interest include child cognitive and socio-emotional development and nutritional status and they were all pre-specified. We measured child development across several key domains including cognitive and language development, receptive language, school readiness, executive functioning and socio-emotional development. We summarise and combine individual measures using exploratory factor analysis to create our main child development outcome measures. We create factor measures that link to three broad areas of child development - (1) cognitive development, language development and school readiness (which we refer to as CLS); (2) pre-literacy skills; and (3) socio-emotional skills. Further, we use four measures of nutritional status - weight for age, length/height for age, BMI for age and weight for length.

Overall we find that the HIM program alone had no impact on child development in the broad CLS domain. We do, however, see evidence of an impact of HIM+FE on this measure corresponding to an increase of 0.15 of a standard deviation (p=0.025), relative to the control group, and find that these effects are substantially larger for older children than for younger children. The difference between the effects of HIM+FE and HIM alone on CLS is statistically significant. We see no significant impact of either HIM or HIM+FE on pre-literacy. We find no evidence that either HIM or HIM+FE had a positive average effect on children's socio-emotional development and, finally, we find no significant impact on any binary measure of nutritional status for the HIM+FE group.

We explore a number of potential mechanisms. Our main finding here relates to impacts of the intervention on measures of the quality of care provided in the centres. The largest and most statistically significant effect is a negative impact of HIM+FE on “Personal Care Routines” of 83 per cent of a standard deviation. The effects on the other three sub-scales are all positive, significant at the 10 per cent level and roughly one half of a standard deviation. These HIM+FE effects are consistent with the core emphasis in the pedagogical training implemented as part of HIM+FE on focusing efforts on productive learning routines and away from non-productive personal care routines.

Several lesson emerge regarding programme design and implementation. First, it is not enough to increase spending or to bring in more people or learning resources: teaching practices and daily experiences of children need to improve. Staffing centres with new professionals without appropriate training and clearly defined roles which are directly linked to improving classroom routines are unlikely to be effective. Second, training programs for early education teachers need to incorporate specific recommendations of ways in which purposeful use of play, art, music and exploration can actually facilitate the child’s learning as translating these ideas into practice can be challenging. The training provided by the HIM+FE intervention, hands-on training with weekly assignments aimed at generating materials to be used for specific activities with children, provides a useful and potentially scalable model for improving the classroom experience of pre-school children.

# Contents

## List of figures and tables

## Abbreviations and acronyms

| | |
|---|---|
| ASQ:SE | Ages and Stages Questionnaires for the socio-emotional domain |
| BMI | Body mass index |
| CLS | Cognitive development, language development and school readiness |
| DCAS | *De Cero a Siempre* (From Zero to Forever) |
| ECD | Early childhood development |
| ECERS-R | Early Childhood Environment Rating Scale **–** Revised |
| FCI | Family Care Indicator |
| FE | Fundación Éxito |
| FTE | Full-time equivalent |
| HI | *Hogare infantile* (preschool) |
| HIM | *Hogares infantiles mejorados* (enhanced preschools) |
| ICBF | Instituto Colombiano de Bienestar Familiar |
| ITERS-R | Infants and Toddlers Environment Rating Scale – Revised |
| IRT | Item response theory |
| ProPAN | Process for the Promotion of Child Feeding |
| PTT | Pencil Tapping Task |
| TVIP | *Test Visual de Imágenes Peabody* (Peabody Picture Vocabulary Test) |
| WHO | World Health Organization |
| WM | Woodcock-Muñoz Tests of Achievement |

# 1. Introduction

It is now widely recognised that the early years are a critical period of a child's development, both because of the extent of their influence on later life outcomes and the extent to which development during these years can be influenced by intervention. Policies targeted at the early years have received a considerable amount of attention recently, with growing recognition of their potential to have long-lasting effects. However, there is still insufficient evidence on the effectiveness and efficacy of different interventions during the first years, and their optimal design. While more and more early childhood interventions are being evaluated, the evidence on specific programmes remains limited, particularly in developing countries. This evaluation makes an important contribution to this evidence.

In its National Development Plan for 2010–2014, *Prosperidad Para Todos*, the government of Colombia announced that early childhood development (ECD) was one of its priorities and launched the *De Cero a Siempre* (DCAS, meaning 'From Zero to Forever') programme as the centrepiece of its social policy. As reported in the National Development Plan, 1.04 million socio-economically vulnerable children under age 6 did not have access to integrated early childhood services, 1.27 million had access to non-integrated services and close to 0.57 million had access to reasonably integrated early childhood services. Hence, one of the main components of this strategy, and the first to be fully articulated and established, is the upgrading of a key non-integrated service – preschools providing partly subsided day care and 60 per cent of daily nutritional requirements to children from low socio-economic backgrounds, at the cost of US$10–25 per month to the parents. These preschools, called *hogares infantiles* (HIs), are run by the Instituto Colombiano de Bienestar Familiar (ICBF); there are 1,008 of them across Colombia, and they are community-based – typically located in fairly well-equipped community centres and run by parent associations.

The upgrade of HIs to *hogares infantiles mejorados* (enhanced HIs, or HIM), implemented as part of DCAS, entailed a considerable financial commitment by the government and was planned to include (1) hiring additional and better-qualified personnel (including nutritionists, experts in child socio-emotional development and pedagogical assistants) and (2) the delivery of a one-time pedagogical endowment for a value of US$52 per child for toys, books and other materials. In addition to the HIM upgrade, some HI centres received a set of further improvements, provided by a private foundation, Fundación Éxito (FE). These improvements (HIM+FE) included three core programmes: further nutritional improvement in HIM centres, pedagogical training of HIM teachers and a reading programme for HIM teachers, children and parents.

The government is committed to evaluating the upgrade of HIs and allowed us to build an evaluation around its expansion. We designed a controlled social experiment, taking advantage of the expansion phase of HI enhancement. Our design allows us to evaluate the effects of the government-implemented improvements only (HIM), as well as of the add-on effects of the enhancements delivered by FE. From 120 HIs, we randomly selected 40 to belong to each of three arms: HI (control), HIM and HIM+FE. We collected data on a sample of 1,989 children (17 per centre on average and 663 per arm) who were enrolled in the selected HIs at the start of the study, their households and the centres, both before the enhancement of HIs (baseline) and 18 months after (follow-up).

The main questions our evaluation seeks to address are:

1. What are the impacts of HIM and HIM+FE, compared to HI, on children's nutritional status and on motor, cognitive, language and socio-emotional development? Does the HIM+FE package have an added impact on these outcomes, relative to HIM?
2. Are the additional resources required by HIM and HIM+FE worthwhile in terms of the returns that they deliver?
3. Do the impacts of the programme vary by child characteristics such as gender, age, race or ethnicity and initial developmental levels?

It is worth noting that, given that the government upgrade of HIs was implemented nationwide, its evaluation is not to be subjected to the common criticism that only a small pilot is being evaluated.

We find that children in HIM+FE centres (those that received both the government and the private foundation improvements) performed significantly better on assessments of cognitive and language development than children attending the control HI centres. The effects of the programme come through most strongly for the older children in the sample across most measures of child development. We do not find significant improvements in measures of child development as the result of the HIM programme. Further, across most measures of child development the HIM+FE programme had significantly more positive effects than the HIM programme. We find limited evidence of impacts of the two programmes on children's nutritional outcomes. The results show that the HIM+FE programme had some positive, marginally significant impact on anthropometric measures linked to weight – weight for age, body mass index (BMI) for age and weight for length z-scores. There is also some preliminary evidence to suggest that the HIM programme have led to a decrease in the proportion of overweight children. Overall, the most persistent, stable key finding emerging from this evaluation is that the HIM+FE improvements were successful in raising children's cognitive and language development, on which the government HIM programme by itself had no impact.

Investigating the impact of HIM+FE impact on children's cognitive and language development further, we find that it holds particularly strongly for boys, children from homes with higher baseline levels of stimulation and children taught by more experienced teachers. These findings suggest that there might be complementarities between the impacts of the HIM+FE programme and stimulating home and school environments. Analysis of intermediate outcomes suggests that a possible mechanism for the effect of HIM+FE on children's cognitive and language development is a change in the types of activities on which teachers spend their time, away from care routines and towards educational play. These changes are consistent with the content of the pedagogical training component of the HIM+FE programme.

The rest of this report is organised as follows: We start by describing the background and context of this study in more detail (Section 2); we then describe the evaluation design and data collected for this study in Section 3. Section 4 discusses qualitative and quantitative evidence on the success of implementation of both programmes. The main estimates of programme impacts and analysis of potential mechanisms are presented in Section 5. We discuss the key policy recommendations emerging from the main findings in Section 6, and conclude in Section 7.

# 2. Background and context

## 2.1 Background

The proportion of children enrolled in early childhood education in Colombia increased from 44 per cent in 2010 to 60 per cent in 2013 (Bernal and Quintero 2014). The Santos administration (2010–2014) decided to prioritise early childhood investment as a result of increased interest in and evidence about the importance of pre-primary education programmes for growth, development and equity. In 2011, the government launched the national early childhood strategy, DCAS. The strategy aimed to deliver high-quality and integrated ECD services for 1.2 million disadvantaged children, with a budget close to US$1,290 million per year, over four years (Bernal and Camacho 2012).

A board of policymakers from various national agencies involved in the provision of early childhood services (including ICBF, the Ministry of Education, the Ministry of Health, the National Planning Department and the Ministry of Culture) was established and has met regularly since mid-2011. This board, Comisión Intersectorial de Primera Infancia*,* was coordinated directly from an office in the presidency and led a major initiative to establish a set of national pedagogical guidelines for early childhood services in Colombia. This document was meant to establish the general goals of the national early childhood strategy rather than specific programmatic guidelines. It discussed the developmental milestones that children should achieve and the conceptual framework that should characterise early childhood services, such as learning based on play, music and art; self-exploration; and opportunities to learn to control emotions and feelings. In addition, the board determined several areas of work aimed at improving quality of services, which included training of care providers; protection of children's rights; inclusion of parents in initiatives related to their children; provision of adequate health and nutrition services; and provision of adequate, safe and inclusive environments for children (Comisión Intersectorial de la Primera Infancia 2012a). DCAS also placed an important emphasis on local capacity building to strengthen the ability of subnational governments to design and operate services at the local level.

## 2.2 HIM

With the above objectives in mind, DCAS has implemented various specific initiatives since 2011. One such initiative, and the first to be fully articulated and established, is the upgrading of HIs to 'enhanced' HIs, or HIM. HIs are medium-sized childcare centres that provide partly subsidised day care and 60 per cent of daily nutritional requirements to children 2–5 years of age from low socio-economic backgrounds. They employ three to six teachers with some training in early education, each caring for about 30 children. Some centres also enrol children below 2 years of age. They are run by the ICBF and are community-based, in the sense that they are typically located in fairly well-equipped community centres and are run by the parent associations.

The upgrading to HIM consisted of two components:

1. Hiring a team of professionals, which includes one health or nutrition professional per 200 children, one expert in socio-emotional development per 200 children and one pedagogical assistant per 50 children to aid with activities in the classroom. The cost of this component is approximately US$20 per child per month.

2. Delivery of a one-time pedagogical endowment for a value of US$52 per child for toys, books and other materials.

These improvements entail a considerable financial commitment: the estimated annual cost of HIM is US$1,330 per child and US$1,000 for HI.[1]

This innovation is fully promoted and financed by the ICBF and is an integral element of DCAS.[2]

## 2.3 HIM+FE

The second intervention we assess is a package of upgrades to the HI programme on top of the HIM improvements. This package was provided by FE, a non-profit organisation in Colombia that aims to improve early education services and, more generally, the well-being of young children in Colombia. FE operates through a variety of strategies, including knowledge generation, direct investments in early childhood programmes as add-ons to existing public programmes, and local capacity building. FE was created in 1982 with an endowment provided by Grupo Éxito, an economic group that owns the biggest chain of supermarkets in the country. From its creation until the year 2000, FE invested in projects aimed at improving social and economic conditions of vulnerable populations in Antioquia (the north-western region of Colombia). From 2000 to 2006, the emphasis shifted to children's nutritional status and education, in addition to health and employment issues. Starting in 2006, FE concentrated all efforts on investments aimed primarily at enhancement of physical, cognitive and socio-emotional development of children in Colombia in early childhood.

For the purpose of this study, we denote the combined HIM and additional upgrades provided by FE as HIM+FE. The additional upgrades consist of three components:

1. A **nutritional improvement programme** that increases calorie provision by 15 per cent over what is provided by a standard HI, so total calorie intake at the HI should be at least 80 per cent of the daily requirement. The nutritional component is also intended to improve the quality of food consumed in the centre. In particular, it aims to increase consumption of fruits, vegetables, protein, iron, zinc and vitamins A and B. The programme is delivered by providing HIs with a monthly food certificate that can be exchanged for food in any store of Grupo Éxito, the main supermarket chain in Colombia. The certificate can be exchanged only for authorised products associated with improvement in nutritional status. Direct food provision is complemented with information and training about healthy habits and nutritional practices provided to both HI personnel and children's parents. The programme emphasises food preparation, buying food, food portions and anthropometric follow-up of children. This content is delivered through multimedia materials that are available online to all beneficiary parents.[3] In order to facilitate implementation of the nutritional component, FE also provides HIs with new anthropometric equipment, and with cookware whenever

[1] These data were provided by the ICBF planning office.

[2] In addition to the elements assessed directly in this study, the improvements to HIs and other early childhood services include elements of institutional capacity and human capital building. However, these components are not part of this evaluation.

[3] http://www.nutrimosvidas.com/

necessary. The total cost of the nutritional improvement programme is US$11 per child per month. FE tracks this and other investments through what it calls the Control Panel. The tool was launched in 2006 and is used to collect data on the nutritional status of children that are served by their programmes. In particular, FE keeps track of children's identification numbers, malnutrition, obesity, overweight, affiliation to the health system, de-worming, nutritional habits and parental participation in activities associated with FE's investments. Based on the information, FE follows up on programmes and implements improvements when necessary.

2. A **training programme** for HI teachers which was planned, coordinated and offered by professionals from the Colombian National University and financed by FE. The programme aims to train care providers in how to design and implement activities that promote infant development across the core main areas. It covered the following topics: (1) technical guidelines for early childhood services defined by DCAS; (2) child development from 18–36 months of age; (3) nutrition; (4) brain development; (5) cognitive development; (6) early literacy; (7) the use of art, music, photography and body language for child development; (8) mathematical concepts during early childhood; and (9) pedagogical strategies during early childhood. The training was delivered using four components: (1) in-person instruction through 16 monthly three-hour sessions delivered via videoconferencing technology by early childhood experts from the Colombian National University departments of psychology and education: (2) video tutoring sessions of three hours per week, in which participants worked with their tutors online on their portfolio of activities or other activities studied during in-person meetings; and (3) onsite coaching, in which instructors carry out one classroom observation of participating teachers to provide specific feedback on their content and pedagogical methodology. The programme is offered free, but participating teachers incur costs of transportation to monthly sessions, require Internet access and need materials for preparation of their portfolio of activities.
3. A **reading programme** for children, teachers and parents. The programme promotes reading amongst children and their families, placing emphasis on reading as a way for parents to bond with their children. It provides parents and teachers with strategies to motivate the children to read. It has three components: reading and music promotion, encouraging effective bonding in families through reading, and building an appropriate environment for reading in the centre. The design and delivery were commissioned to Fundación Ratón de Biblioteca, a Colombian NGO devoted to promoting good reading habits amongst youth. The programme starts with a diagnostic stage assessing reading materials, reading areas and opportunities for reading for and with children in participating HIs. Based on the results of the diagnostic stage, a plan and timeline are determined for each HI. The programme provides 150 books per participating HI, and a 'reading promoter' is assigned to serve the HI. This person plans fortnightly workshops with children with appropriate activities around books, stories and fantasy. Likewise, parents attend monthly workshops in which they receive training on how to approach children's books and how to promote their children's reading habits. Finally, the programme also delivers workshops for teachers of children between 12 and 36 months of age to introduce them to techniques on how to promote reading for non-readers. Workshops for parents and teachers

take place at the HI, and teachers' sessions are organised in groups by city. Finally, in addition to the 150 books per HI, 50 book bags are distributed in each treated HI. Book bags are intended to be shared by children in classrooms and taken home. The total cost of FE's innovation is US$22 per child per month.

The total expenditure of FE on the three components of the programme throughout the duration of the intervention was of US$1.2 million; 62 per cent of this expenditure was on the nutritional component alone. A total of 7,100 children were served through the various programmes provided by FE.

## 2.4 This study

The Colombian government has been committed to the evaluation of the upgrade of HIs since its inception, and hence allowed us to design a large-scale cluster randomised controlled trial around the expansion of these enhancement interventions in order to assess the short-term impacts of the two different but complementary upgrades to HIs on child development outcomes. Members of the research team have been in close communication with ICBF and other agencies in the Colombia government responsible for ECD throughout the research study to ensure that the evaluation answers the questions of key interest to them.

The study took place in the eight largest cities in Colombia, which also had the largest number of HIs: Bogotá, Cali, Medellín, Barranquilla, Bello, Palmira, Itagüí and Soledad. These are shown on the map in Figure 1. For logistical purposes, we limited the study to these eight cities, as adding more would incur a huge additional operational cost in return for very few additional HIs interviewed. The selected cities constitute a reasonably varied group. They represent at least four regions in the country – Andean (centre), Atlantic coast, Pacific coast, and south-eastern region – and are very diverse in their economic conditions, cultural traditions and race and ethnicity. This should provide heterogeneity in the study sample and increase the relevance of our findings for different regions of the country. In this study, we aim to test whether the programmes (HIM and HIM+FE) reached their main objectives – improving child growth and development through improvements to the quality of childcare services. In order to achieve this aim, we test three groups of hypotheses:

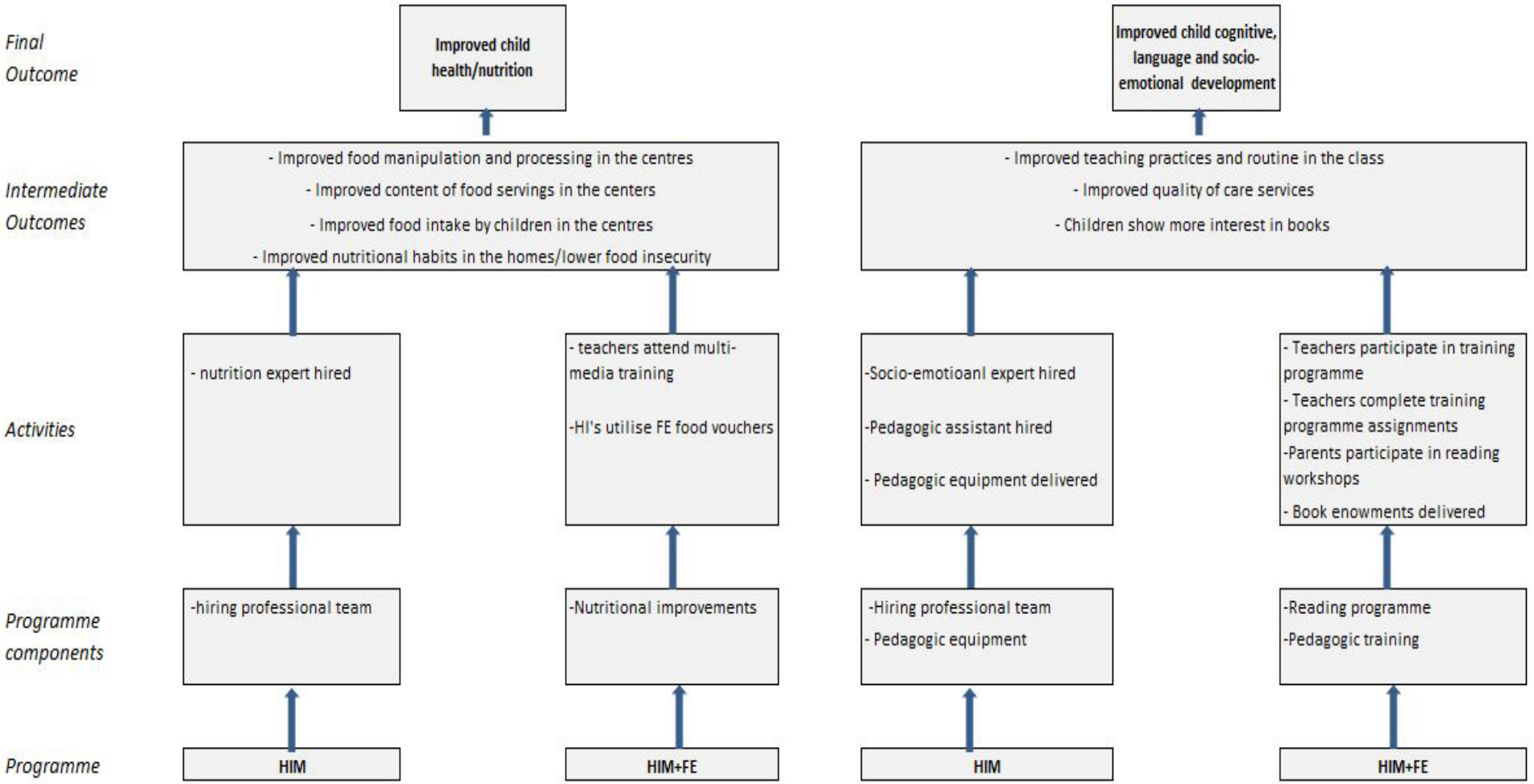
- **Impact on children's outcomes:** HIM and HIM+FE may have positive average impacts on outcomes for children attending the centres. We group these outcomes into two areas: children's development and children's nutritional status.
- **Mechanisms:** HIM and HIM+FE are more likely to improve children's outcomes if the improvements they aim to introduce in the centres are delivered effectively.
- **Heterogeneity of impacts:** The individual characteristics of children, their teachers and the HI centres may determine the extent to which children benefit from HIM and HIM+FE.

**Figure 1: Study locations**


The complete set of hypotheses is set out in the pre-analysis plan and can be found in Appendix C. Figure 2 demonstrates our hypothesised mechanisms for programme impacts separately for child development and nutritional status outcomes; each programme component is linked to the activities involved (inputs and upgrades) which, in turn, link to the intermediate outcomes that we hypothesise may constitute the key mechanisms by which the interventions affect final outcomes. In the remainder of this report, we present findings linked to each level of Figure 2.

In Section 4, we describe in detail activities linked to each component of the HIM and HIM+FE programmes, as well as results from analysis of qualitative and quantitative programme implementation data. We present the main estimates of programme impacts on final outcomes in Section 5.3 and explore heterogeneity in these effects in Section 5.4. In the last part of this report (Section 5.5), we focus on the impact of the programme on intermediate outcomes, which, according to our hypotheses (set out above and in Figure 2), constitute the key mechanisms linking programme activities with targeted final outcomes.

**Figure 2: Mechanisms**



# 3. Evaluation

## 3.1 Evaluation design

### *3.1.1 Objectives of the evaluation*

In this evaluation, we aim to determine whether the enhancement of HIs proposed by ICBF (HIM) is effective and cost-efficient. We also wish to ascertain whether the add-on intervention implemented by FE generates substantial positive effects on child development. Ethical clearance for the research was granted by two review boards: the Ethics Committee for Research at Universidad de Los Andes and the Research Ethics Committee at University College London.

### *3.1.2 Study design*

We use a cluster randomised controlled trial design to estimate the impacts of the HIM and HIM+FE programmes. The randomised design allows us to be confident that the impacts we estimate are indeed causal, i.e. that they are fully attributable to the programme and are not the result of pre-existing differences between different HIs. We use a three-armed (HI [control], HIM and HIM+FE) design to ensure that we are separately able to estimate three effects: (1) the effect of the HIM upgrades to HIs relative to HIs without any upgrades, (2) the combined effect of the HIM+FE upgrades relative to HIs without any upgrades and (3) the effect of the HIM+FE upgrades relative to only the HIM upgrade. For each effect, we are able to use statistical methods to test whether each effect is significantly positive, i.e. that we can be at least 95 per cent confident that a positive effect we estimate arose through the true effect being positive, rather than by chance. This corresponds to testing whether there is evidence to believe that (1) the HIM upgrades have a positive effect on outcomes relative to HI, (2) the HIM+FE upgrades have a positive effect on outcomes relative to HI, and (3) the HIM+FE upgrades have a more positive effect on outcomes than the HIM upgrades alone.

Because the intervention we are evaluating is implemented at the HI level, random assignment into the two treatment arms and the control group was also done at the HI (cluster) level. However, we measure impacts on children's outcomes by comparing children registered in HI that were enhanced or that received the enhancement plus the FE add-on intervention to children registered at unchanged HIs. All analysis is performed on an intention-to-treat basis, i.e. we look at the impacts of the two programmes on children registered at the HIs at the start of our study, regardless of whether they attended the centre throughout the intervention period.

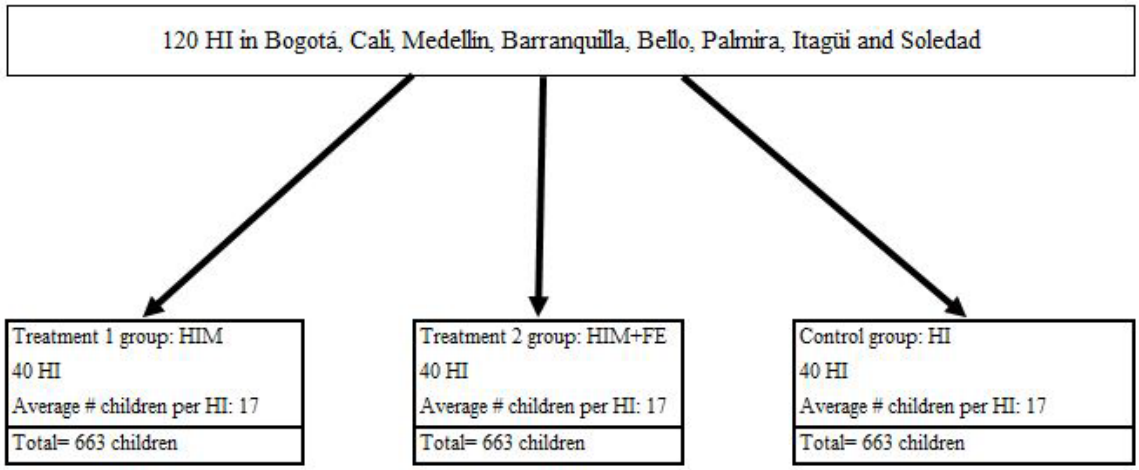
### *3.1.3 Sample size*

We used power calculations to determine the sample size, both in terms of the number of HIs and the number of children in HIs, that we would require to be able to detect moderate positive effects of the two programmes on two key outcomes, if they were indeed present, with a probability of 80 per cent. For this purpose, we considered effects of the programmes on height for age and cognitive development, and we defined moderate positive effects as 20 per cent of one standard deviation for HIM as compared to HI and 20 per cent of one standard deviation for HIM+FE as compared to HIM. Because all of the children in our sample are treated by one modality or the other – i.e. they received some sort of early childhood education – the scope for improvement is lower than if we had an untreated control group who received no early childhood

education. For the power calculations, we assumed a moderate intra-cluster correlation co-efficient, which was similar to what we observed in the baseline data for cognitive development (0.035).[4]

Given our assumptions, we calculated that we would require 15 children per cluster (HI) to achieve 80 per cent power at a 5 per cent significance level. To allow for attrition between baseline and follow-up data collection (of about 10%), we assessed 17 (whenever there were enough) children per cluster. This led to a target sample size for baseline of 2,000 children in 120 clusters (since some HIs had only 15 or 16 children in the target age range).

**Figure 3: Programme evaluation design**




### *3.1.4 Sampling and randomisation*

From November–December 2012, we randomly selected 198 HIs from 670 in our eight study cities. These 198 HIs were then organised into groups of three geographically close HIs. From these groups, we selected 40 for inclusion in the study, based on information on attendance by age range. In particular, we needed to guarantee that each HI in our sample had at least 15 children in our target age range of between 12 and 36 months at baseline. Then, within each group of three, we randomly assigned one HI to the control group, one HI to treatment group 1 (HIM) and one HI to treatment group 2 (HIM+FE). We chose to stratify our randomisation using these groups of three to increase the likelihood that our sample was balanced, i.e. comparable on baseline characteristics, so that any differences observed between the groups at follow-up could be attributed to the programmes. A cluster (HI) had, on average, 48 children between 12 and 36 months of age from which the baseline sample of 17 children per HI was drawn.

### *3.1.5 Attrition and balance*

Our attrition rate was relatively low; only 125 (6.29%) of the 1,987 children in the baseline sample were not found in the follow-up round. Table 1 shows that attrition rates were similar between treatment groups, and Appendix Table B.14 confirms that attrition was not significantly related to treatment assignment. In Appendix Table B.15 we review balance at baseline over key child and household characteristics. As discussed in the baseline report, we find that the sample is largely balanced, with the exception of the

[4] As measured by the Ages and Stages Questionnaire.

HIM group's having a significantly higher proportion of male children and lower weight for age z-scores than the control group. When we recreate the same table for the sample who we retained at follow-up (in Appendix Table B.16), these two imbalances remain, but no others emerge. This confirms that attrition was not related to treatment status, which could have biased our results. We address the two baseline imbalances by controlling for children's gender and baseline values of anthropometric scores in relevant analysis.

**Table 1: Sample size and attrition by treatment status**

| | Control | HIM | HIM+FE | Total |
|---|---|---|---|---|
| Sample at baseline | 661 | 663 | 663 | 1,987 |
| Sample at follow-up | 616 | 617 | 629 | 1,862 |
| Attrition rate | 6.81% | 6.94% | 5.13% | 6.29% |

## 3.2 Data

### *3.2.1 Data sources*

We collected data to inform on impact of the programmes on (1) final outcomes of the children in the treatment and control groups (child cognitive, language and socio-emotional development; health and nutritional status); (2) intermediate outcomes (teaching practices and routines in the class, quality of care services, and food manipulation practices, content and intake at the centres); (3) activities directly related to the implementation of the programmes; and (4) other factors that can affect child development, including characteristics of other children in the family and parents, household circumstances, characteristics of the HIs that they attend.

The questionnaires administered during baseline and endline data collection include:

- **Household:** This questionnaire collected information about the socio-demographic characteristics of the child's household, details about the father, mother, head of the household (in case it is neither the child's father or mother) and the main caregiver (in case it is the father, mother or head of household).
- **Child in household:** This questionnaire collected specific information about the child who attends the HI. It contained specific questions about the history of participation in HIs and other forms of childcare and his or her health and nutritional status. These were answered by the main caregiver, ideally, or the adult who was available at the time of the survey.
- **HI:** We collected information that includes a physical description of the HI and compliance with minimum quality standards (ICBF guidelines). After inspecting the HI's infrastructure, furniture, spaces and other details, the assessor indicated whether ICBF quality standards were met. The guidelines include items related to hygiene, sanitation, infrastructure, personnel and other areas. We also collected information about the teachers and classrooms, and collected specific information from the HI's director, pedagogical assistant and personnel for socio-emotional support (in HIs where one was present). Additionally, we conducted a separate expert assessment of quality of care provision at a subset of HIs using the Early Childhood Environment Rating Scale assessment instrument, discussed below.

- **Child in HI:** We collected information about the classroom and teacher, as well as some relevant information about the child in the HI. We also collected all available administrative information about the child.
- **Consumption:** We measured food consumption for children in the HI in a subsample of 50 HIs. This is of particular importance since improved nutrition is a feature in HIM+FE. We verified the menus by weighing the portions. We also collected information from the HI's nutritionist (in HIs where one was present).

The full list of questionnaires and the survey round in which they were administered is presented in Appendix Table A.13.

We now describe how key intermediate and final outcomes were measured.

***Measures of final outcomes***

The main developmental outcomes examined in this study link to a range of domains, including some aspects of cognition, language, school readiness, pre-literacy skills, socio-emotional development and some aspects of executive functioning. We measured functioning in these domains using several tests. We were careful to select tests that have been used extensively in evaluations of early care or education or have been recommended for developing countries (Fernald et al. 2009). These measures have demonstrated adequate psychometric properties in similar longitudinal studies with Latin American or other Hispanic populations and have effectively captured programme effects; most of them have been used in other ECD studies in Colombia (Bernal and Quintero 2014; Bernal and Fernández 2013). Finally, members of the research team have prior experience in the use of most of these instruments in Colombia. The final selection of the scales was made in consultation with a child development psychologist.

- **Cognitive and language development.** The Woodcock-Muñoz III Tests of Achievement (WM-III) is a comprehensive set of individually administered tests of children's early literacy and mathematical skills and knowledge. We use subtests 5, 12, 14, 17 and 21 to measure concept formation, fluid reasoning, expressive language, memory for words and rhymes (Muñoz-Sandoval et al. 2005). Following guidelines for conducting this assessment, subtest 17 (memory for words) was administered to all children older than 36 months at follow-up (all but one child) and subtest 21 (rhymes) was administered to a subsample of children older than 42 months at follow-up (75% of the full sample). The scales have been translated into Spanish, adapted for Latin American contexts and used in previous studies to evaluate effects of early childhood interventions on cognitive development in infants and older children (Fernald et al. 2009).
- **Receptive language.** *Test Visual de Imágenes Peabody* (TVIP) (Dunn et al. 1986), the Spanish version of the Peabody Picture Vocabulary Test, is a standardised measure of receptive language. It has been used extensively with Spanish-speaking populations in the United States and in Latin America (Schady et al. 2015), correlates well with individual IQ scores (Stanford-Binet Intelligence Scale, $\rho = 0.62$) and serves as a good predictor of school readiness and later academic achievement.
- **School readiness.** The Daberon-2 test provides a standardised assessment to screen for school readiness (Danzer et al. 1991). The test covers topics related to body parts, recognising colours and numbers, gross motor development, categorisation, recognising and using prepositions and other abilities related to

early academic success. It is administered individually and takes between 20 and 40 minutes. We did not use the full list of items in the original test. Instead, we selected a subset of items that were most suitable for the age range of children in our study sample. We then piloted our selected items to ensure that there was sufficient variation in test performance amongst our target population.

- **Executive functioning.** We use the Pencil Tapping Task (PTT), which requires the ability to exercise inhibitory control over one's prepotent behaviour, the natural tendency to mimic what the experimenter does. In particular, it requires holding two things in mind: (1) to tap once when the experimenter taps twice and (2) to tap twice when experimenter taps once (Diamond and Taylor 1996). It thus measures inhibitory control and working memory. In line with the guidelines for conducting this assessment, it was only administered to children over the age of 48 months at the time of follow-up (60 per cent of the full sample).
- **Socio-emotional development.** We used the Ages and Stages Questionnaires for the socio-emotional domain (ASQ:SE) (Squires et al. 2002b) for all children in baseline and follow-up. The ASQ:SE is a parent-completed assessment system for children ages 6–60 months, completed through culturally sensitive questionnaires focusing on socio-emotional development and the identification of children at risk of social-emotional difficulties. It measures self-regulation, compliance, communication, adaptive functioning, autonomy, affect and interactions with others. The ASQ:SE shows high levels of consistency, reliability, validity and specificity (Squires et al. 2002a, 2002b) and has been used for early development assessments in low and middle low income countries (Handal et al. 2007; Heo and Squires 2012). To reduce the impact of literacy, ASQ:SE was collected through parent interviews.

Our other key set of outcomes relates to child nutritional status. In line with similar international studies (e.g. Fernald et al. 2008; Walker et al.), we collected information on height and weight, BMI and arm circumference, following World Health Organization (WHO) standards for all children in our sample at baseline and follow-up (WHO Multicentre Growth Reference Study Group 2006). Data collectors assessed growth according to standard protocols after undergoing training provided by professional nutritionists and achieving the required level of inter-observer reliability in their anthropometric measurement.

For all developmental tests, we report impacts on age-standardised scores and scores by subdomain. We use the anthropometric measures to construct a variety of nutritional indicators depending on the child's age and based on WHO (2006) standards. We present an in-depth discussion of standardisation procedures applied to test scores and anthropometric measures in Section 5, below.

***Measures of intermediate outcomes***

As shown in Figure 2, we expect the programme to have an impact on child development and nutrition outcomes through improving the quality of the children's home and preschool environments. We utilise a number of existing, internationally validated scales to capture features of these environments considered crucial for child development:

- **Early Childhood Environment Rating Scale – Revised (ECERS-R)** (Harms et al. 1998) provides a global measure of preschool classroom quality, with 43 items that cover a broad range of quality dimensions, including safety, teacher-child

interaction and parental involvement. This measure has been used extensively and has well-established validity and reliability in a wide range of countries and cultural and economic contexts. The ECERS-R has shown predictive validity to child gains across cognitive domains (Peisner-Feinberg et al. 2001; Burchinal et al. 2000) and social-emotional domains (Sylva et al. 2006).

- **Infants and Toddlers Environment Rating Scale – Revised (ITERS-R)** (Harms et al. 2003) provides a global measure of environmental quality for infants and toddlers with 39 items that cover a broad range of quality considerations similar to the ECERS-R, and it also has been widely used in the field internationally.
- **Quality of the home environment.** We collected the number of play materials (toys the child usually plays with) and play activities the child engaged in with an adult over the seven days before the interview using the items included in UNICEF's Family Care Indicator (FCI) (Frongillo et al. 2003). The FCI is a short and easy-to-administer test derived from the much longer Home Observations for Measurement of the Environment (Caldwell and Bradley 2003), a well-known measure of the level of home stimuli. The FCI has been validated against Home Observations for Measurement of the Environment in a poor environment (Hamadani et al. 2010); these authors also assessed how well it correlates with the Bayley Scales of Infant Development II (Bayley 1993). We extended the questions on play activities to include some additional activities and to collect the time devoted to each activity. In addition, we asked the caregiver these questions separately.

Both the HIM and HIM+FE programmes contain core components linked to child nutrition. Key measures of the quality of children's nutritional intake include:

- **Process for the Promotion of Child Feeding (ProPAN):** This is a tool developed by UNICEF and the Pan American Health Organization to develop, implement and assess interventions and programmes to improve diet and eating habits for infants and toddlers. The ProPAN software returns calorie intake by food groups (protein, iron, calcium, vitamins A and C and zinc) by meal and by age range. In order to utilise this tool, we measured food consumption for children in a subsample of 50 HIs. The methodology consists of a visit on a random date and interviews with kitchen personnel, together with weighing and observation of two randomly picked morning snack servings and two randomly picked lunch servings. Each of these is split into servings for children younger than 2 years and servings for children older than 2 years, if different.
- **Standards of food manipulation and processing:** This is captured by measuring compliance with ICBF guidelines in a number of domains – people involved in food preparation, food preparation area, food storage area, area where food is consumed and practices around mealtimes.

This is not intended to be an exhaustive list of measures, but rather an indicative list of the key measures linked to the main intermediate outcomes presented in Figure 2.

### *3.2.2 Data collection procedures and quality control*

Parental permission to participate in the study and parent questionnaires (on the household's socio-demographic characteristics and children's socio-emotional outcomes) were collected in HIs whenever possible. Otherwise, the assessment took place in the household. Background information on teachers and centre staff was

collected directly from them. Data on child's attendance status at follow-up were collected directly from school records available to the evaluation team and double-checked using parental questionnaires.

The ECERS-R and ITERS-R were completed by psychologists who observed centre classrooms for at least half of a school day. Assessors were trained to be unobtrusive, to the extent possible, in the classroom while observing and recording. The team of four psychologists was trained for three weeks and assessed for reliability in two centres in Bogotá, which were not in the study sample. Child outcome data were collected by psychologists who were trained and assessed for reliability by the evaluation team. Children were assessed in the HI in a designated area with an adequate environment for the child (e.g. the director's office or the nursing area). Children not attending centres were assessed in their own homes. In all cases, assessors were instructed to guarantee a quiet and comfortable space for these activities.

During fieldwork, data collection was carried out by six teams. Each team consisted of a supervisor, an assistant to the supervisor, four assessors and two psychologists in charge of direct child assessments. The assistant to the supervisor was in charge of interviewing professional centre staff and service providers and checking for compliance with administrative guidelines by direct observation of centres. Assessors were in charge of collecting parental socio-demographic characteristics and child socio-emotional development by parental report, and administering teacher questionnaires. Assessors were trained for three weeks. The training included a module on measurement and standardisation of anthropometric measures for precise and accurate information. The training was provided by a professionally certified nutritionist with vast expertise in measurement and training. It also included training on each of the evaluation instruments and interviewing techniques.

Prior to fieldwork, the supervisor contacted the HI director to inform her or him about the visit and the activities that would take place during the visit. Upon arrival at the HI, the supervisor checked enrolment lists to establish attendance and check which study children were still enrolled at follow-up. All HI evaluation instruments were then administered, including assessment of compliance with guidelines, professional staff, HI director, teachers, child in HI and classroom. Study children in the HI were then assessed in a special area designated for this purpose. If the child did not attend the HI, then he or she was assessed in his or her own home. Families were provided with a monetary incentive of US$12 once the assessment was completed, in the form of a gift card to the biggest supermarket chain in the country.

Data collection was computer-based in most cases, with the exception of classroom observations and child assessment. Each supervisor centralised the digital information in her computer and transmitted it to the headquarters in Bogotá every night. Paperwork, including signed parental consent forms, was shipped once a week to headquarters. All of the information was tracked electronically by a control map designed for the study. The quality control software generated periodic indicators that allowed tracking of coverage, assessor efficiency and completeness of the child files. This allowed timely corrections or appropriate adjustments to work in the field. Classroom observations and child assessment data were entered in the headquarters in Bogotá. Digital data were

stored securely on a server at headquarters and fully compiled into data tables by the end of fieldwork.

Various types of supervision were implemented during fieldwork:

- Direct supervision: a supervisor worked along with the assessor (with minimal interference with the assessor's work). Direct observation guaranteed that the information registered in the system was true.
- Indirect data verification:
  - Phone calls: random phone calls to households, centres or teachers were made and selected information was confirmed.
  - Re-interviews: the supervisor occasionally re-visited the households, centres or teachers to confirm selected information.
- Coaching of fieldwork personnel: continuous support provided by the group supervisor ensured that data quality was maintained.

Once the complete data set was received by the research team, it was assessed for consistency and completeness. We made sure that all data sets could be merged and that identification numbers were complete and correct. We also verified that total interviews reported by our fieldwork provider (by unit, such as by centre, child, household or teacher) coincided with the actual data. Minor inconsistencies or missing data were reported to our fieldwork provider. In some cases, they returned to the field or contacted centres, households or individuals by phone to complete missing or inconsistent information. This iteration process took about one month. After this time, the research team finalised data processing and initiated the statistical analysis.

Baseline data collection on the children, their households and the centres they attended (including surveys to centre staff) were collected between March and June 2013, with the exception of six centres in Barranquilla, for which data could not be collected until August 2013 (because of resistance from the teacher's union to participate in the study). Follow-up data collection, initially planned 12 months after baseline, took place after 15–19 months, during October and November 2014. The decision to delay follow-up data collection was motivated by a series of implementation delays (see next section).

## 4. Programme implementation

The randomisation was performed in November and December 2012, followed by the baseline (March–June 2013), after which both the upgrades (HIM and FE) were due to begin. However, there were two important delays to this process. First, while the FE upgrade was introduced immediately after the completion of baseline data collection in June 2013, its teacher-training component was not completed until May 2014. Second, while the roll-out (implemented by ICBF) began in February 2013, HIs required a substantial amount of time to hire their complete interdisciplinary teams of professionals (nutritionists and health professionals, experts in socio-emotional development and pedagogical assistants). Because of these delays, the research team, FE and ICBF made the decision to delay follow-up data collection, originally planned for March–June 2014, until October–November 2014. It was agreed that this delay would ensure that there was enough time for the upgrades to take effect before impact evaluation.

We now discuss implementation of each programme component in detail.

## 4.1 HIM

### *4.1.1 Hiring and roles of the new personnel*

The HI budget increase associated with the HIM upgrade was allocated in February 2013 (about a month before the start of baseline); HIs were directly responsible for hiring the additional personnel (nutritional experts, socio-emotional experts and pedagogical assistants). No penalties were incurred for not complying with the programme. In other words, treated HIs received the full budget increase regardless of how many professionals they had hired. ICBF drafted the job description for each of the additional personnel to be hired. The profile and description of tasks is included in Appendix A.

Overall, the hiring picture is mixed. Ninety per cent of HIM centres had the two professionals and pedagogic assistant employed at the time of follow-up.[5] However, many of these professionals worked part time and so, when we look at full-time equivalent (FTE) professional-to-child ratios, we see a very different picture, with far fewer of the HIM centres meeting the required ratios. Overall, only five of the 80 HIM centres (6.3%) met the required ratio for the complete set of two professionals and pedagogic assistant (Appendix table E.23).

Given the design of HIM, we suggest three possible explanations for why we do not observe full compliance. First, there could have been too few people with the required qualifications in the local area. Second, since HIM centres received the funds for hiring new personnel regardless of whether they actually hired them, centres may have spent the money on alternative centre needs (e.g. materials, infrastructure or wages for other staff). Finally, the money could have been used for expenditures unrelated to centre needs.

The data we have do not allow us to identify how the HIM-related HI budget increase was used. We have some tentative evidence, especially from the qualitative evaluation data, that local labour market conditions, including labour supply and wages, constrained hiring of the new professionals. In the quantitative data, while we do not see large differences in compliance rates between the eight cities in the study-sample (Appendix Table E24), in the larger cities there is some evidence of clustering, with centres in some areas of the city appearing more likely to have met the hiring requirements.[6] We also see cases (two for socio-emotional experts and nine for nutritional experts) in which the same professional was employed by multiple (usually two) HIMs. Though not conclusive, these patterns could be indicative of local shortages in the availability of suitably qualified candidates.

Despite partial compliance, the overall picture from the qualitative evaluation interviews and focus groups with parents, centre directors and teachers suggests that there was a high degree of awareness of the responsibilities of the new personnel and general impression that they contribute to building and maintaining the relationship with parents

[5] Interestingly, centres in the control group had also hired some personnel; this might have been because they had heard about the HIM upgrades to centres nearby and decided to also invest additional resources into hiring. However, the scale of hiring in the control group was significantly smaller than HIM for all three categories of personnel.

[6] We do not include maps showing this pattern in order to preserve the anonymity of the centres in the sample.

and provision of high-quality care in the centres. However, it also reveals that a large proportion of the time of all three types of personnel was spent in administrative duties and paperwork; they also often worked in isolation (rather than being integrated into the wider team) and were restricted in their ability to carry out their tasks by lack of office space. Finally, the additional personnel were reportedly sometimes used to cover the absences of other staff members rather than to perform the roles for which they were hired.

Below, we describe hiring figures within each category (nutritional expert, socio-emotional expert, pedagogic assistant) and evidence on their realised roles (extracted from the qualitative evaluation). Because there do not appear to be any significant differences in the hiring of HIM professionals between the HIM and HIM+FE groups (Appendix Table E.23), we simply describe the combined hiring figures.

***Nutrition expert***

Twenty months after baseline, at the time of follow-up data collection, all but one of the centres assigned to receive the HIM upgrades had hired at least one nutritional expert, compared with one in five of the controls. Before adjusting for part-time work, we find that three-quarters of HIM centres had at least one nutritional expert present per 200 children (the target). However, many of the nutritional experts worked part time and when we adjust for this, we find that only one in eight had at least one FTE nutritionist per 200 children (compared with one in 40 controls).

In the qualitative evaluation, nutritionists reported their main tasks as administration of all processes related to food provided to children in centres, revision of menus, supervision of portions served, monitoring of special cases, anthropometric follow-up, training kitchen personnel and working with families on better nutrition and health-related habits. The qualitative evaluation revealed that many, particularly those who had been hired part time, felt that they had little time for the amount of work required. In particular, they often reported that administrative tasks and paperwork took up too much of their time.

***Socio-emotional expert***

A similar story emerges in relation to the socio-emotional experts. All of the HIM centres had at least one socio-emotional expert employed at the time of follow-up (compared with 37.5% of the controls). However, again, many worked part time, so that just 30 per cent of HIM centres fulfilled the requirement of one FTE socio-emotional expert per 200 children (compared with 5% of the controls).

During the qualitative interviews, socio-emotional experts reported their main tasks as support to parents on all issues within the socio-emotional domain, handling of domestic violence, diagnostic assessments and preparation of plans for cases requiring follow-up (e.g. ones in which developmental lags were identified). It was also reported that an important component of their job was their work with families, which relieved teachers from some of the responsibilities and was generally perceived as allowing teachers to spend more time and energy directly in the classroom. Like the nutritional experts, the socio-emotional experts indicated that they had too little time for the amount of work and had too many administrative duties. There were cases where socio-emotional experts were being used to cover absences of other staff.

***Pedagogic assistant***

All but one of the 80 HIM centres had at least one pedagogic assistant employed at the time of follow-up, compared with 15 per cent of the control centres. Most HIM centres had more than one pedagogic assistant, with the average being 3.44 FTE. Overall, though, only 38.8 per cent of HIM centres met the requirement of one FTE pedagogic assistant per 50 children, compared with none of the controls.

The qualitative evaluation revealed that pedagogic assistants reported their roles as support for personal care routines in the nursery (particularly of younger children), replacing absent teachers, helping at lunch and snack times and helping with administrative duties and paperwork, especially records of children's progress and planning of activities The general perception is that their help with the administrative duties relieved teachers of tasks they had previously carried out in the evenings. Pedagogic assistants were generally assigned to classes of younger children; therefore, classes of older children were still often found to be led by a single teacher.

## 4.2 HIM+FE

### *4.2.1 Nutritional improvements*

The implementation components targeting nutrition in the HIM and HIM+FE programmes (described in Section 2) were assessed through an apparent consumption evaluation carried out in a random subsample of 50 HIs in Bogotá, Cali, Medellín and Barranquilla, including centres in the three study groups. Data collection and processing were led by a professional nutritionist. During each meal (morning snack, lunch and afternoon snack) three random portions by age range (18–23 months, 24–36 months, older than 36 months) were weighed at each centre. This information was complemented with reports by the kitchen personnel on preparations and ingredients. In addition, apparent consumption was assessed through direct observation by the nutritionist during mealtimes to register whether children consumed most of the food, three-quarters, half or one-quarter of the portion served. This was done based on previously agreed and standardised criteria designed by the team of nutritionists at FE.

Calculation of nutritional intake was carried out using ProPAN software. The nutritional assessment included only the quantities actually consumed by (not served to) the child. Using this information, we calculated the rate of nutritional compliance for each age group and for each of the nutritional groups analysed. Nutritional compliance is calculated as the amount consumed of a given nutritional group, divided by the requirement established by the centres' nutritional guidelines (depending on whether it is ICBF's guidelines in HIM or FE's guidelines in HIM+FE).

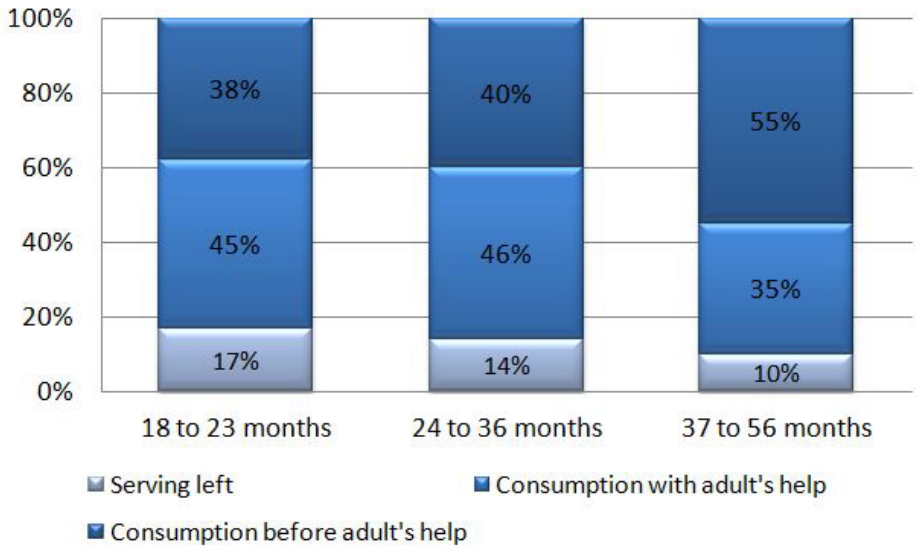
The following are the main results from the nutritional evaluation:

1. The percentage of weight of served portion, with respect to weights established in the nutritional guidelines, is 84.5 per cent for children 18–23 months of age, 91.5 per cent for children 24–36 months of age and 84.7 per cent for children older than 36 months in HIM centres. These compliance levels are higher in the HIM+FE group, with 99.7 per cent for children 18–23 months of age, 101 per cent for children 24–36 months of age and 93.1 per cent for children older than 36 months. Finally, compliance rates in the control group were 102.3 per cent for

children 18–23 months of age, 102.3 per cent for children 24–36 months of age and 104 per cent for children older than 36 months.

2. In HIM, approximately 35 per cent of children report consumption of nutritional supplements in their household. In the case of HIM+FE, this fraction is about 27.9 per cent and 33.5 per cent in the control group.
3. With respect to changes in children's menus in cases of nutritional risks or other chronic diseases, such as diabetes, we report that 31 per cent of centres in HIM implement an adjustment in cases of low weight, 31 per cent in cases of overweight and 31 per cent in cases of diseases. These percentages are 47 percent, 29.4 percent and 47 per cent in HIM+FE centres, respectively. Finally, in the control group, adjustments for any of the above risks were made in about 35 per cent of cases.
4. During mealtimes, we found that children in the 18–23-month age range consume approximately 38 per cent of their serving on their own and an additional 45 per cent with the help of the pedagogical assistant, and do not consume 17 per cent. For children 24–36 months of age, close to 40 per cent is consumed by the child on his or her own and an additional 46 per cent with the help of the pedagogical assistant or the teacher, and approximately 14 per cent is not eaten. Finally, in the group of children older than 36 months, we observe that 55 per cent is consumed by the child on his or her own, an additional 35 per cent with the help of an adult and 10 per cent not at all (Figure 4).

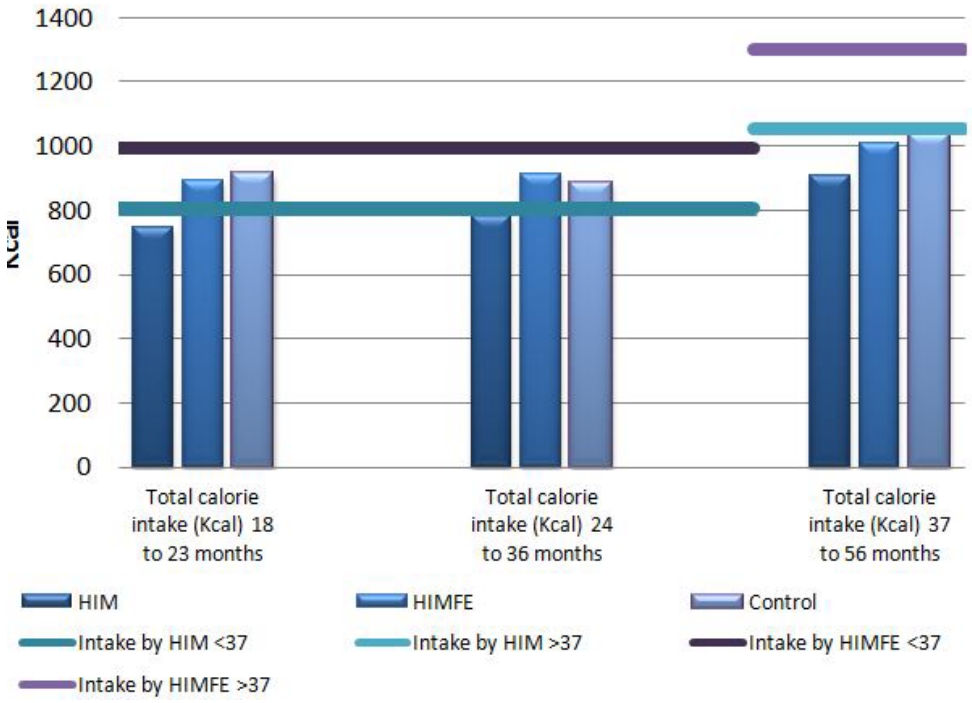
**Figure 4: Total calorie intake**




In Figure 5, we report total calorie intake (actual consumption) and compare it with what it is supposed to be, given nutritional guidelines (ICBF and FE). ICBF guidelines for children younger than 36 months of age are 807 kilocalories, and for children older than 36 months are 1,052 kilocalories. In the case of FE guidelines for HIM+FE centres, the figures are 992 kilocalories for children younger than 36 months and 1,299 kilocalories for children older than 36 months. As can be observed in Figure 5, compliance is slightly lower in HIM with respect to ICBF guidelines – in particular, 747 kilocalories for children 18–23 months of age, 782 for children 24–36 months of age and 906 for children older than 36 months. In HIM+FE centres, total calorie intake is higher than ICBF guidelines

but never reaches levels determined by FE guidelines – e.g. 891 kilocalories for children 18–23 months of age and 910 kilocalories for children 24–36 months of age (versus 992 kilocalories), and 1,010 kilocalories for children older than 3 months (versus 1,299). Finally, in the control group we observe that total calorie intake is higher than ICBF guidelines for children younger than 36 months of age and exactly what it should be for children older than 36 months; it is 914 kilocalories for children 18–23 months of age and 885 kilocalories for children 24–36 months of age.

**Figure 5: Total calorie intake (actual consumption)**




Finally, consumption of carbohydrates was 61 per cent in HIM and 59 per cent, on average, in the control group, while this fraction was 54 per cent in HIM+FE. These percentages are similar across age groups. Although these rates fall within tolerable adult minimum daily requirement levels, consumption of simple carbohydrates is higher than 10 per cent in all age groups and treatment status, which implies higher risks of overweight and obesity.

### *4.2.2 Reading programme*

The reading component of the HIM+FE programme, described in Section 2.3, was implemented between June 2013 and December 2014 (three semesters). Final reports by Fundación Ratón de Biblioteca indicate that 2,592 children were reached; 6,000 books (150 per HI) and 2,000 book bags (50 per HI) were delivered; and 3,637 workshops for children, 328 workshops for parents and 89 workshops for teachers were held. At follow-up, directors of centres in the HIM+FE group reported that their centres had 72.8 more books on average than centres in the control group (Appendix Table E.25). FE personnel also monitored the progress of the programme collecting indicators, including numbers of children who were interested in books, asked parents to read to

them, read books with images, recognised letters in books and expressed ideas about the story in the book during classroom reading sessions.[7]

The reading promoters reported that implementation of the FE reading component was hindered by lack of appropriate physical space for reading activities within HIs (e.g. too much noise, poor illumination and lack of equipment for digital books and music). They also reported that participating children changed from 2013 to 2014 because the programme targeted children from 12–36 months of age and a significant fraction of children outgrew this age range during the programme. For that reason, implementation had to be adapted to start with new groups of beneficiaries every academic period. All book endowments were successfully delivered, but the participation of parents in workshops was low. Our qualitative evaluation revealed that parents are, in general, aware of the importance of reading for the development of their children but lack appropriate books at home, do not find time to read to their children or report feeling unprepared for the task, and, in some cases, indicate that young children will not understand much anyway.

#### *4.2.3 Pedagogic training*

The pedagogic training for teachers, detailed in Section 2.3, was implemented between June 2013 and June 2014. Given the delivery method of the intervention, teachers needed access to computers and the Internet in order to participate. Centre directors nominated two or three teachers per treated HI to participate, with some additional teachers from the same centres selected to replace teachers who were not able to attend all of the sessions or dropped out. Administrative records show that in all, 114 of the teachers in the 40 HIs assigned to HIM+FE started the training, a figure in line with what FE had planned. Of these, 99 were certified as having completed it – a completion rate of 87 per cent. Although the training was designed for teachers, other staff were often sent, including teaching assistants and directors or other senior staff.

## 5. Impact analysis

In this section, we present analysis of the impact of the HIM and HIM+FE programmes on the main child developmental and nutrition outcomes. We start with a discussion of the empirical strategy and our main outcome measures of child development. We then present and discuss estimates of programme impacts. In the remainder of the section, we explore these impacts through heterogeneity analysis, as well as impacts of the programme on intermediate outcomes, which we hypothesise constitute the mechanisms through which the programmes could have an impact on targeted outcomes (Figure 2).

Throughout this section, we closely follow the pre-analysis plan, using the empirical strategy and indicators specified there and testing the full set of hypotheses (Appendix C).

[7] It is important to note that the programme ran concurrently with a national initiative aimed at promoting reading in early childhood services nationwide, known as *Fiesta de la Lectura* (Reading Party). This means that HIM and control HIs were also being served by this initiative. The strategy consisted mainly of delivery of an endowment of age-appropriate books, which were largely received by centres before the baseline of this study. It did not involve any substantial elements of training, coaching or parental involvement.

## 5.1 Empirical strategy

Given the experimental design described in Section 3, we can identify the impacts of HIM (Treatment 1, or $T_1$ ) and HIM+FE (Treatment 2, or $T_2$) on final outcomes as:

$$Y_{icsl,1} = \beta_0 + \beta_1 T_{1sl} + \beta_2 T_{2sl} + \gamma Y_{icsl,0} + X'_{icsl,0}\theta + C'_{icsl,0}k + Z'_{icsl,1}\lambda + \epsilon_{icsl,1}$$

where $Y_{icsl,1}$ is the outcome of interest for individual $i$ in class $c$ in childcare centre $s$ in city $l$ at follow-up (t=1); $T_{1sl}$ is a dummy equal to 1 if the childcare centre $s$ in city $l$ receives Treatment 1 (HIM); $T_{2sl}$ is a dummy equal to 1 if the childcare centre $s$ in city $l$ receives Treatment 2 (HIM+FE); and $Y_{icsl,0}$, if applicable, is the baseline (t=0) level of the outcome of interest for individual $i$ in class $c$ in childcare centre $s$ in city $l$. For child development outcomes, we do not have the same measure at baseline and follow-up, since we administer different tests given children's different ages in each data collection round. For these outcomes, we control for the key baseline scores relating to child development (ASQ – broken down by sub-scale – and MacArthur-Bates Communication Developmental Inventory). The purpose of including these baseline measures as controls is to maximise efficiency. $X_{icsl,0}$ is a set of basic child characteristics (age and sex), which are also added to improve efficiency (minimise residual variance) and control for the slight imbalance in the proportion of male children whom we observed between groups at baseline; $C_{icsl,0}$ are a complete set of city dummies and $Z_{icsl,1}$ are a complete set of tester or interviewer dummies, depending on whether the outcome was measured by the tester (psychologist) or the household interviewer. $\epsilon_{icsl,1}$ is the random error term, clustered at the childcare centre levels (the unit of randomisation).

When the outcome of interest is continuous, we estimate equation 1 by ordinary least squares. $\beta_1$ is the estimated average impact of HIM on outcome $Y_{icsl,0}$ and $\beta_2$is the estimated average impact of HIM+FE on the outcome (both are intent-to-treat estimates). In this case we report the mean of the control group.[8] We then report the differences ($\beta_1$, $\beta_2$ and $\beta_2 - \beta_1$ ), their corresponding standard errors and the p-value associated with testing the null hypothesis that controlling for baseline outcome measures, child age and sex, as well as city and tester effects, differences in outcome measures at the end of the intervention between HIM and control, HIM+FE and controls and HIM+FE and HIM are zero.

Where the outcome of interest is binary (e.g. stunting) we estimate equation 1 using a logit model.[9] In this case, we report the estimated adjusted difference in proportion of the outcome in question between the treatment groups and the standard error of that difference (calculated using Stata's margins command). We do not use as dependent or independent variables any indicator variable that has a prevalence rate of below 10 per cent or above 90 per cent, in order to limit noise caused by variables with minimal variation.

[8] As long as the variable is not standardised internally, in which case the mean is zero by construction and we omit it from the tables.

[9] While this was not specified in the pre-analysis plan, it is generally recognised as a better way of estimating effects for binary outcome data.

For most children's developmental outcomes, we use one-tailed hypothesis tests. This is motivated by a strong prior that the upgrades would not harm children's developmental levels. For a few of the final outcomes (e.g. children's weight) and all of the intermediate outcomes, we use a two-tailed test, since we do not have as strong a prior about the effects of the two interventions. We exclude from the analysis children with developmental outcomes or nutritional status with standardised values lower than three standard deviations below the mean (<-3SD) of the relevant standardised distribution, since we consider this to be an indication of potential disability (for developmental outcomes) or severe malnutrition (for nutritional status).

We impute missing covariate values (for $Y_{icsl,0}$ and $X_{icsl,0}$), with the average of the non-missing observations and account for the imputation with a dummy variable. In cases for which the percentage of observations with covariate missing data is less than 2 per cent, we simply work with the subsample with non-missing data.

Finally, where appropriate, we adjust p-values for multiple hypothesis testing. This is important to assure us that any significant results we find are not simply the result of testing for an effect of the intervention(s) on multiple outcomes and focusing on ones that are significant. We report both corrected and uncorrected p-values for all estimates of impacts on children's cognitive skills, socio-emotional skills and nutritional status. We use two methods to correct our inference for multiple testing. First, we use the Romano-Wolf step-down procedure on our multiple measures of child cognitive development and school readiness (Romano and Wolf 2005a, 2005b). This is a more powerful method than other options, since it allows for interdependence between hypotheses. We use the Romano-Wolf method to correct our inference on our multiple measures of child cognitive development and school readiness. However, we run into problems when using this method on outcomes collected by the interviewers who visited the home (anthropometry and socio-emotional skills) because of the presence of many fixed effects, which causes problems in the bootstrap procedure. For these outcomes, we use the Šidák procedure to correct for multiple hypothesis testing (Šidák 1967), which is substantially more conservative than the Romano-Wolf method, since it assumes the hypotheses to be independent. Given how conservative the Šidák correction is, some marginally significant estimates (based on Šidák-corrected p-values) could still be interesting to consider.

## 5.2 Measures of child outcomes

### *5.2.1 Child developmental outcomes*

We analyse the impact of HIM and HIM+FE on a number of domains of child development (discussed in Section 3.2). In this section, we first discuss how we construct scores for each individual measure of child development and then describe how these are used to construct the main child development outcome measures for the impact analysis.

#### *Individual measures of child development*

In total, we used nine instruments to measure aspects of cognition, language, school readiness, pre-literacy skills, socio-emotional development and executive functioning. The instruments used and the dimensions of child development they measure are shown in Table 2. As specified in the pre-analysis plan, all assessments were scored in line with

the instructions in the relevant administration manuals. In order to remove the effects of age and increase the efficiency of the outcome measures test, developers also provide assessment-specific standardisation procedures. However, these rely on norms in populations that are unlikely to be comparable to ours. In the main analysis we therefore use internal standardisation procedures (i.e. using test data collected for the study sample) to construct the outcome measures for the individual assessments.[10]

We use item response theory (IRT) to construct outcome measures for six of our eight assessments of cognitive and language development: WM5, WM14, WM17, WM21, DABERON and TVIP (as indicated in Table 2). In the context of this study, using IRT has two advantages. First, because it explicitly allows assessment items to vary by how difficult they are and how well they discriminate between children of different underlying abilities, its estimates of underlying abilities are more efficient than an approach that relies on aggregation of the raw scores (if the assumptions are satisfied). Second, several of our tests (in the WM batteries) have non-continuous distributions of raw scores (Figure D.8) created by particular stopping rules (e.g. if child scores six points or fewer of the first available nine, then stop assessment; otherwise continue). With assessments that use such rules, different children attempt different questions, so it is not appropriate to simply use raw scores calculated by adding up the number of correct answers; IRT provides a way of constructing valid outcome measures (see discussion in Appendix D for more details).

The two remaining assessments are not suitable for IRT: WM12 is simply in the form of a count variable, and there is little reason to believe that different items are more or less discriminatory than others in the PTT. For these two assessments, we generate a score by adding up the number of correct responses. The outcome measures for all these assessments, as well as the ASQ raw score of socio-emotional development, are then re-scaled to have a mean of zero and standard deviation of one for the control group. Distribution plots for each assessment, along with a more detailed discussion, are in Appendix D.

***Measures of child developmental outcomes***

As set out in the pre-analysis plan, since child development is composed of many interrelated dimensions, we would expect there to be strong correlation between the different child development measures. We therefore combine the individual measures (constructed as described above) using exploratory factor analysis to estimate common underlying constructs that we use as our main child development outcome measures.[11] This improves the efficiency of the way in which we use the assessment data and avoids problems linked to testing multiple outcomes for a single hypothesis. We distinguish between assessments that link to three broad areas of child development: (1) cognitive development, language development and school readiness (which we refer to as CLS), (2) pre-literacy skills and (3) socio-emotional skills. The ASQ:SE is the only instrument used to measure socio-emotional development. Table 2 shows which assessments were included in factor analysis conducted to identify the most appropriate way of combining

[10] For completeness, we also provide impact estimates on outcome measures standardised using the external, tester-provided norms in Appendix Table E.26, discussed below.

[11] Further discussion and tables showing correlations in test scores are in Appendix D.

assessments in constructs within each of the other two areas of child development. CLS includes receptive language (TVIP), concept formation (WM5), fluid reasoning (WM12), expressive language (WM14), memory for words (WM17), school readiness (DABERON) and inhibitory control (PTT). Pre-literacy skills include receptive language, expressive language, memory for words and rhymes (WM21).[12] Both for CLS and pre-literacy skills, factor analysis yields one factor that complies with our inclusion criterion of having an eigen value greater than one. As indicated in Table 2, two of the assessments had factor loading of less than 0.4 and were therefore, as per the analysis plan, excluded from the constructed factors (WM5 for CLS factor and WM21 for the pre-literacy factor). The constructed factors were re-scaled to have a mean of zero and standard deviation of one for the control group. More details about the construction of the two-factor measures, including tables showing the loadings of the individual assessments onto the two factors, as well as distributions of these, can be found in Appendix D.

An issue worth noting is that following testing protocols, the test of inhibitory control (PTT) was administered only for children over the age of 48 months at the time of follow-up, i.e. roughly 60 per cent of the sample. As a consequence, the factor for CLS constructed using the most complete set of tests is defined for only these 60 per cent of the sample. We construct another factor for CLS that excludes PTT and is therefore defined for children of all ages.

In line with the pre-analysis plan, in addition to the main analysis of impacts on the factor measures, we also report impacts on individual assessments, including p-values corrected for multiple hypothesis testing using the Romano-Wolf step-down procedure (Romano and Wolf 2005a, 2005b), as discussed in Section 5.1.

[12] Note that three out of four assessments that make up pre-literacy are also included in CLS.

**Table 2: Child development instruments**

| | Instrument used | Dimension of child development measured | Outcome measure constructed using IRT | Included in CLS factor | Included in pre-literacy skills factor |
|---|---|---|---|---|---|
| TVIP | Test de Vocabulario en Imágenes de Peabody (Dunn et al. 1986) | Receptive language | Yes | Yes | Yes |
| WM5 | Batería-III Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 5 (Muñoz-Sandoval et al. 2005; Woodcock 1977) | Concept formation (cognition) | Yes | * | |
| WM12 | Batería-III Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 12 (Muñoz-Sandoval et al. 2005; Woodcock 1977) | Fluid reasoning (cognition) | | Yes | |
| WM14 | Batería-III Woodcock-Muñoz: Pruebas de Aprovechamiento – 14 (Muñoz-Sandoval et al. 2005; Woodcock 1977) | Expressive language | Yes | Yes | Yes |
| WM17 | Batería-III Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 17 (Muñoz-Sandoval et al. 2005; Woodcock 1977) | Memory for words (cognition) | Yes | Yes | Yes |
| WM21 | Batería-III Woodcock-Muñoz: Pruebas de Aprovechamiento – 21 (Muñoz-Sandoval et al. 2005; Woodcock 1977) | Rhymes | Yes | | * |
| DAB | Daberon-II Screening for School Readiness (Danzer et al. 1991) | School readiness | Yes | Yes | |
| PTT | Pencil Tapping Task (Diamond and Taylor 1996) | Inhibitory control | | Yes | |
| ASQ:SE | Ages and Stages: Socio-Emotional Questionnaire (Squires et al. 2002) | Socio-emotional | | | |

Note: * Initially included in factor (as set out in pre-analysis plan) but excluded due to low loading (<0.4).

### *5.2.2 Child nutrition outcomes*

We use four measures of nutritional status – weight for age, length or height for age, BMI for age and weight for length. These scores are standardised relative to a WHO reference population of healthy children. The mean z-score in the control group is –0.399 for weight for age and –0.795 on length or height for age. This indicates that the average child in our population (without the effects of HIM or HIM+FE) is around 40 per cent of a standard deviation lighter and 80 per cent of a standard deviation shorter than the average child in a healthy population. Strong correlations for all four measures between baseline and follow-up imply that, as expected, these measures are very persistent.

In addition to z-score measures of nutritional status, we also estimate programme impacts on a range of binary indicators of nutritional status (listed and defined in Table 3, below).[13]

We only consider the impact for indicators that have between 10 per cent and 90 per cent prevalence in our sample as a whole.

**Table 3: Nutritional status indicators**

| Indicator | Definition |
|---|---|
| Chronic malnutrition | Length or height for age z-score < –2 |
| Low height | Length or height for age z-score ≥ –2 and < –1 |
| Adequate height | Length or height for age z-score ≥ –1 and ≤ 1 |
| Risk of low weight for height | Weight for length or height z-score ≥ –2 and < –1 |
| Adequate weight for height | Weight for length or height z-score ≥– 1 and ≤ 1 |
| Overweight | Weight for length or height z-score > 1 and ≤ 2 |
| Overweight (BMI) | BMI z-score > 1 and ≤ 2 |

## 5.3 Impact on children's outcomes

The first of our three hypothesis groups, as set out in the pre-analysis plan (Section 2.4 and Appendix C), is that both HIM and HIM+FE may have positive average impacts on outcomes for children attending the centres. We group these outcomes into children's developmental outcomes and children's nutritional status.

### *5.3.1 Child cognitive and language development, school readiness and pre-literacy skills*

As we show in Figure 2, we hypothesise that HIM may improve CLS and pre-literacy skills through hiring more professionals and support for the centres, as well as the pedagogic equipment. Additional components of HIM+FE that could affect these outcomes include the reading programme and pedagogic training. In order to examine whether each of the two programmes had a positive average impact on these outcomes, we test the mean difference between each treatment group and the control group after

[13] A key reason for doing this is that it can be difficult to interpret estimated programme impacts on nutritional status by simply using z-scores as measures of nutrition outcomes. For example, larger effect on z-scores linked to weight is not unambiguously better in contexts, like this one, where there is a prevalence of overweight children in the population.

controlling for baseline developmental levels, child age and sex, as well as tester and city effects (see specification in Section 5.1).

***Cognitive development, language development and school readiness***

As discussed above, the full set of test scores used to construct the CLS measure is only available for the older 60 per cent of the sample, as PTT cannot be administered to children younger than 48 months of age. In order to estimate impacts of HIM and HIM+FE on CLS for the whole sample (including the younger children), an outcomes measure that excludes PTT has to be used. We proceed in the following way: Row 1 in Table 4 shows estimates of impacts on the older sample for whom the complete set of test scores is available for construction of the CLS measure. Row 2 shows estimates for the same older sample as Row 1 but using the CLS measure that can be constructed for the whole sample, i.e. excluding PTT. Finally, Row 3 shows estimates of impacts on the same measure as Row 2, but for the whole sample, and Row 4 shows estimates of impacts of HIM and HIM+FE on pre-literacy skills.

Table 4 shows that the HIM programme alone had no impact on child development in the broad CLS domain (p=0.663). We do, however, see evidence of an impact of HIM+FE on this measure (p=0.025) corresponding to an increase of 0.15 of a standard deviation, relative to the control group. The effect persists for this subsample when using the alternative measure (available for the full sample) that excludes scores for PTT. The effect disappears once the age range is extended to include the younger children. This pattern of results suggests that the impacts of HIM+FE are substantially larger for older children in the sample than for younger children; we investigate this issue further in Section 5.4.

***Pre-literacy***

The last row of Table 4 shows estimates of programme impacts on the factor measure of pre-literacy skills. We see no evidence of a significantly positive impact of HIM or HIM+FE on this outcome (p=0.925 and p=0.249, respectively). Indeed, if we were using a two-tailed t-test, we would interpret the estimate as suggesting that the HIM programme (but not HIM+FE) had adverse effects on pre-literacy.

***Discussion***

We investigate these results further by estimating programme impacts on standardised values of individual test-scores (constructed as described in Section 5.2.1) that make up the measures of child development used in Table 4. Understanding programme impacts on these individual components will offer insights into the underlying trends driving the main results.[14]

We find that HIM programme participants scored less well, on average, on TVIP and WM14 (expressive language) than children in the control group (Table 5). HIM+FE had positive impact on WM12, which measures fluid reasoning.

The components that make up the CLS measure include all of the tests listed in Table 5. Estimates in Table 5 suggest that positive overall effect of the HIM+FE programme are

[14] Appendix Table E.27 further shows programme impacts on each component of the Daberon instrument. Just as with the overall DABERON score, we find no impacts on any of the sub-scales.

driven by positive impacts on WM12 and, more weakly, by effects on WM14 and the PTT.

The pre-literacy measure is made up of TVIP, WM14 (expressive language) and WM17 (memory for words). Estimates in Table 5 suggest that the negative HIM programme coefficient in Table 4 is to a large extent driven by the poor performance of HIM participants on TVIP.

Exploring the estimates further, we compare the effect sizes of the HIM and HIM+FE programmes. Tables 4 and 5 show evidence that supports our hypothesis in the pre-analysis plan that the effect of HIM+FE on outcomes will be at least as large as the effect of HIM. For both specifications of the CLS factor, as well as the factor relating to pre-literacy skills, we comfortably reject the null hypothesis that the impact of HIM+FE is equal to the impact of HIM alone using a two-tailed t-test ($p=0.007$, $p=0.025$ and $p=0.013$, respectively; Table 4).[15]

In Table 5 we see a similar pattern, with the impact of HIM+FE significantly greater than that of HIM alone for WM14 and marginally so for MW 12.

[15] The comparison between the HIM+FE and HIM groups is starker than the comparison between the HIM+FE group and the control because the point estimates of the mean HIM only effects on the three factors are actually negative (although not significantly different from zero).

**Table 4: Impact of HIM and HIM+FE on (1) cognitive development, language development and school readiness factor(s) and (2) pre-literacy factor**

| | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value2 | Adjusted difference HIM+FE-HIM | p-value3 | N |
|---|---|---|---|---|---|---|---|
| Cog, Lang and Sch (all measures, limited sample) | –0.030 | 0.663 | 0.151** | 0.025 | 0.182*** | 0.007 | 1,071 |
| | (0.072) | | (0.076) | | (0.073) | | |
| Cog, Lang and Sch (exc. PTT, limited sample) | –0.041 | 0.712 | 0.148** | 0.028 | 0.189*** | 0.006 | 1,071 |
| | (0.073) | | (0.076) | | (0.074) | | |
| Cog, Lang and Sch (exc. PTT, full sample) | –0.066 | 0.828 | 0.066 | 0.162 | 0.132** | 0.025 | 1,819 |
| | (0.069) | | (0.067) | | (0.066) | | |
| Pre-literacy skills | –0.100 | 0.925 | 0.045 | 0.249 | 0.145** | 0.013 | 1,819 |
| | (0.069) | | (0.066) | | (0.064) | | |

Notes: Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is greater than the control group mean (one-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the child (a total of 1,987 children's households were interviewed). All estimates control for gender, city effects, tester effects and baseline scores for MacArthur-Bates Communicative Development Inventories and each sub-scale of the ASQ-III. All factors scaled to have a mean of zero and standard deviation of one in the control group. All factors constructed as described in section 5.2.1. IRT scores used to construct factors (where applicable). Age effects removed from standardised scores prior to factor construction. Each factor constructed using the following standardised scores: (1) Cog, Lang and Sch (all measures): WM12, WM14, WM17, TVIP, DAB and PTT; (2) Cog, Lang and Sch (exc. PTT): WM12, WM17, TVIP, WM14 and DAB; and (3) pre-literacy skills: WM14, WM17 and TVIP.

**Table 5: Impact of HIM and HIM+FE on all child developmental and pre-literacy measures**

| | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | N |
|---|---|---|---|---|---|---|---|
| TVIP | –0.084 | 0.936 | –0.011 | 0.754 | 0.073 | 0.393 | 1,819 |
| | (0.066) | [0.896] | (0.065) | [0.570] | (0.062) | [0.122] | |
| Woodcock Muñoz – 12 | 0.006 | 0.734 | 0.148*** | 0.009 | 0.142 | 0.102 | 1,825 |
| | (0.057) | [0.458] | (0.059) | [0.007] | (0.068) | [0.019] | |
| Woodcock Muñoz – 14 | –0.084 | 0.977 | 0.099 | 0.220 | 0.183** | 0.013 | 1,825 |
| | (0.070) | [0.884] | (0.065) | [0.064] | (0.062) | [0.002] | |
| Woodcock Muñoz – 17 | –0.072 | 0.878 | 0.021 | 0.496 | 0.093 | 0.188 | 1,825 |
| | (0.058) | [0.891] | (0.060) | [0.364] | (0.058) | [0.056] | |
| Pencil Tapping Task | 0.038 | 0.702 | 0.100 | 0.268 | 0.061 | 0.227 | 1,073 |
| | (0.074) | [0.302] | (0.076) | [0.097] | (0.073) | [0.200] | |
| DABERON – total | –0.003 | 0.827 | 0.055 | 0.514 | 0.058 | 0.338 | 1,825 |
| | (0.068) | [0.515] | (0.063) | [0.189] | (0.064) | [0.183] | |

Notes: Standard errors in parentheses below estimated treatment effects. P-values adjusted for multiple hypothesis testing using the Romano-Wolf step-down procedure, unadjusted analytical p-values in parentheses below. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is greater than the control group mean (one-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the child (a total of 1,987 children's households were interviewed). All scores internally standardised to remove the effect of age and to have a mean of zero and standard deviation of one in the control group. IRT used wherever applicable. Age effects removed from standardised scores. All estimates control for gender, city effects, tester effects and baseline scores for MacArthur-Bates Communicative Development Inventories and each sub-scale of the ASQ-III. Instruments as designed to measure the following dimensions of child development: TVIP=receptive language, WM12=fluid reasoning (cognition), WM14=expressive language, WM17=memory for words (cognition), DABERON=school readiness, Pencil Tapping Task=inhibitory control. Variation in N across tests is due to missing values for some tests (unintentional) and to PTT only being administered to the older subsample (48+ months at follow-up).

### *5.3.2 Child socio-emotional development*

Table 6 shows that we find no evidence that either HIM or HIM+FE had a positive average effect on children's socio-emotional development, as measured by the ASQ:SE. The estimated impacts on the standardised total ASQ:SE score are small and not significantly different from zero. Comparing the impacts of HIM and HIM+FE on socio-emotional development, in line with hypothesis in the pre-analysis plan, we find no evidence to suggest that HIM+FE had any detrimental effect relative to HIM alone.

**Table 6: Impact of HIM and HIM+FE on socio-emotional development as measured by the ASQ:SE**

| | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | N |
|---|---|---|---|---|---|---|---|
| Self-regulation | 0.075 | 0.543 | 0.062 | 0.772 | –0.013 | 0.985 | 1,796 |
| | (0.056) | [0.093] | (0.064) | [0.169] | (0.054) | [0.594] | |
| Compliance | –0.084 | 1.000 | –0.009 | 0.999 | 0.074 | 1.000 | 1,797 |
| | (0.063) | [0.906] | (0.062) | [0.560] | (0.066) | [0.130] | |
| Communication | 0.099 | 0.257 | –0.013 | 0.984 | –0.113 | 0.271 | 1,791 |
| | (0.055) | [0.037] | (0.056) | [0.595] | (0.063) | [0.961] | |
| Adaptive functioning | 0.014 | 0.986 | 0.024 | 0.968 | 0.010 | 0.990 | 1,796 |
| | (0.065) | [0.415] | (0.062) | [0.350] | (0.063) | [0.438] | |
| Autonomy | 0.068 | 0.542 | 0.030 | 0.916 | –0.038 | 0.903 | 1,797 |
| | (0.051) | [0.093] | (0.048) | [0.266] | (0.056) | [0.747] | |
| Affect | –0.030 | 1.000 | –0.052 | 1.000 | –0.022 | 0.999 | 1,788 |
| | (0.071) | [0.665] | (0.080) | [0.741] | (0.076) | [0.613] | |
| Interaction with people | 0.029 | 0.951 | 0.013 | 0.986 | –0.016 | 0.983 | 1,794 |
| | (0.060) | [0.314] | (0.060) | [0.413] | (0.063) | [0.599] | |
| Total | 0.070 | 0.617 | 0.038 | 0.929 | –0.032 | 0.945 | 1,796 |
| | (0.058) | [0.113] | (0.066) | [0.282] | (0.062) | [0.695] | |

Notes: Standard errors in parentheses below estimated treatment effects. P-values adjusted for multiple hypothesis testing using the Šidák procedure, unadjusted analytical p-values in parentheses below. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is different from the control group mean (one-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the child (a total of 1,987 children's households were interviewed). All scores are internally standardised to remove the effect of age and to have a mean of zero and standard deviation of one in the control group. All estimates control for age, gender, city effects, interviewer effects and baseline scores for the ASQ:SE sub-scale in question. Variation in N across sub-scales is due to missing values for some items in each sub-scale.

### *5.3.3 Child nutritional status*

In this section, we examine impacts of HIM and HIM+FE programmes on nutritional status of children. As per Figure 2, we hypothesise that the HIM programme might improve children's nutrition through the newly hired nutritional expert and HIM+FE through the additional provision of nutritional information, multimedia training for the teachers and FE food vouchers. Table 7 shows the estimated impacts on the z-scores of weight for age, length or height for age, BMI for age and weight for length. We find no evidence of impact of HIM alone on any of the four measures of nutritional status; all point estimates are precisely estimated but very close to zero. There is some evidence of a positive impact of HIM+FE on measures relating to weight. This is in line with expectations, since children's weight responds much faster to changes in diet than height, so is more likely to have been affected by the nutritional component of FE. Based on the Šidák p-values, only the impact on weight for age z-score is statistically significant (p=0.065). However, estimates of impacts on BMI for age and weight for length z-scores are also positive and close to statistically significant (especially given that, as discussed in Section 5.1, the Šidák p-values are likely to be overly conservative in this case).

Table 8 shows no significant impact on any binary measure of nutritional status for the HIM+FE group. This reassures us that the increase in weight we observed in the HIM+FE group was not associated with an increase in overweight children, which would have constituted a detrimental effect of the intervention. Despite not seeing any impact on average weight-based z-scores for the HIM group, we do see significant change at the upper tail of the distribution: there is a fall in proportion of overweight children by about 7 percentage points (as measured by both weight for age and BMI for age).

We now turn to comparison of HIM and HIM+FE impacts. We find evidence that HIM+FE had a significantly larger impact on nutritional status than HIM for two of the four z-score measures of nutritional status (BMI for age and weight for length; Table 7). Table 8 further shows that a significantly higher proportion of children in the HIM+FE group were overweight (by both measures) at follow-up than in the HIM group. However, in both groups (HIM and HIM+FE) the proportion of overweight children was smaller than in the control group.

**Table 7: Nutritional status – z-scores**

| | Control mean | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | N |
|---|---|---|---|---|---|---|---|---|
| Weight for age z-score | –0.399 | 0.034<br>(0.045) | 0.945<br>[0.450] | 0.109*<br>(0.048) | 0.065<br>[0.024] | 0.074<br>(0.042) | 0.161<br>[0.082] | 1,835 |
| Length or height for age z-score | –0.795 | 0.030<br>(0.029) | 0.803<br>[0.302] | 0.016<br>(0.036) | 0.993<br>[0.659] | –0.014<br>(0.034) | 0.992<br>[0.676] | 1,834 |
| BMI for age z-score | 0.154 | –0.016<br>(0.054) | 0.996<br>[0.771] | 0.109<br>(0.058) | 0.192<br>[0.062] | 0.125*<br>(0.056) | 0.088<br>[0.028] | 1,834 |
| Weight for length z-score | 0.117 | –0.008<br>(0.055) | 1.000<br>[0.887] | 0.125<br>(0.059) | 0.114<br>[0.036] | 0.133*<br>(0.056) | 0.066<br>[0.020] | 1,834 |

Notes: Standard errors in parentheses below estimated treatment effects. P-values adjusted for multiple hypothesis testing using the Šidák procedure, unadjusted analytical p-values in parentheses below. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is different from the control group mean (two-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the child (a total of 1,987 children's households were interviewed). Z-score calculated using WHO (2005) growth standards. All estimates control for age, gender, city effects, interviewer effects and baseline z-scores for the measure in question. Variation in N is due to missing values for some measures.

**Table 8: Nutritional indicators**

| | Control proportion | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | N |
|---|---|---|---|---|---|---|---|---|
| Chronic malnutrition | 0.112 | –0.007 | 1.000 | –0.016 | 0.783 | –0.009 | 0.995 | 1,834 |
| | | (0.014) | [0.615] | (0.012) | [0.174] | (0.013) | [0.487] | |
| Low height | 0.300 | 0.012 | 0.988 | 0.008 | 0.996 | –0.004 | 1.000 | 1,834 |
| | | (0.015) | [0.427] | (0.011) | [0.496] | (0.013) | [0.770] | |
| Adequate height | 0.568 | –0.005 | 1.000 | –0.002 | 1.000 | 0.003 | 1.000 | 1,834 |
| | | (0.020) | [0.806] | (0.022) | [0.920] | (0.021) | [0.900] | |
| Adequate weight | 0.616 | 0.015 | 0.993 | 0.007 | 1.000 | –0.008 | 1.000 | 1,835 |
| | | (0.021) | [0.458] | (0.019) | [0.706] | (0.021) | [0.704] | |
| Risk of low weight for height | 0.097 | 0.020 | 0.911 | 0.013 | 0.995 | –0.007 | 1.000 | 1,834 |
| | | (0.018) | [0.261] | (0.019) | [0.479] | (0.019) | [0.706] | |
| Adequate weight for height | 0.690 | 0.040 | 0.629 | –0.013 | 0.999 | –0.053 | 0.110 | 1,834 |
| | | (0.026) | [0.117] | (0.025) | [0.602] | (0.021) | [0.014] | |
| Overweight | 0.156 | –0.072*** | 0.001 | –0.018 | 0.972 | 0.054** | 0.029 | 1,834 |
| | | (0.017) | [0.000] | (0.020) | [0.360] | (0.018) | [0.004] | |
| Overweight (BMI) | 0.167 | –0.068*** | 0.001 | –0.024 | 0.891 | 0.045* | 0.093 | 1,834 |
| | | (0.017) | [0.000] | (0.020) | [0.242] | (0.018) | [0.012] | |

Notes: Standard errors in parentheses below estimated treatment effects. P-values adjusted for multiple hypothesis testing using the Šidák procedure, unadjusted analytical p-values in parentheses below. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is different from the control group mean (two-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the child (a total of 1987 children's households were interviewed). All estimates control for age, gender, city effects, interviewer effects and baseline values of the indicator in question. Variation in N is due to missing values for some indicators.

## 5.4 Programme impact heterogeneity

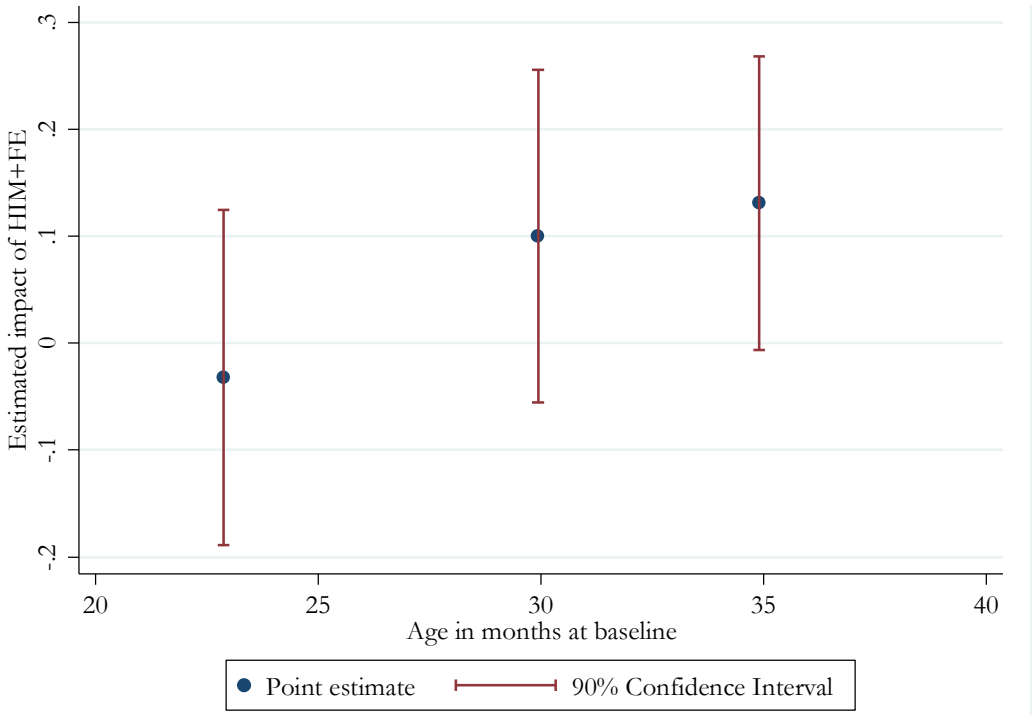
In line with the hypotheses outlined in Section 2.4, we now investigate whether the effects of HIM+FE and HIM on the main developmental and nutrition outcomes vary depending on a selection of child, teacher and HI characteristics detailed in the pre-analysis plan (see Appendix C). The child characteristics include child age and sex, as well as baseline levels of child development, maternal education and level of stimulation in the home. Characteristics of the teacher to whom the child was assigned at baseline include education, experience, baseline teaching practices and routines, depression score, job satisfaction and job 'burnout' score. Finally, HI characteristics include baseline size and ICBF's quality indicators. When the characteristic is continuous (e.g. nursery quality scores or child's age), we split the sample equally around the median. We focus on heterogeneity in the subset of programme effects that we find to be significant for the whole sample in the analysis above.

### *5.4.1 Heterogeneity of programme impacts by child characteristics*

#### *Age*

We state in the pre-analysis plan that we expect to find larger programme impacts for older children than for younger children because most of the tests we administered have higher discriminating power for older than younger children. This is consistent with findings we report above; in Section 5.3.1, we show significant impact of HIM+FE on the CLS factor for an older subsample, but not when the younger subsample was also included (see Table 4). Figure 6 shows coefficient plots of the estimated impact of HIM+FE on this factor (CLS excluding PTT), splitting the sample into three equally sized groups based on age. The confidence intervals displayed are at the 90 per cent level. This coefficient plot further confirms that the effect of HIM+FE on CLS is driven entirely by the children who were in the older part of the sample at baseline. This could be either because the intervention was genuinely more effective for older children or because our tests had higher discriminatory power for this subsample. Appendix Table E.28 shows this heterogeneity broken down by each instrument used to measure child development. For four of the six instruments (WM12, WM14, WM17 and DABERON), the estimated impact of HIM+FE is significantly greater than zero at the 5 per cent level for the older half of the sample (by age at baseline), with only one marginally significant estimated impact for the younger half of the sample (PTT). This pattern also holds for the pre-literacy measure: estimated impact of HIM+FE on the older subsample (older than 30 months at baseline) is substantially larger than for the younger subsample (0.116 versus –0.020) and significant at the 10 per cent level (p-0.059).

**Figure 6: Estimated impact of HIM+FE by age at baseline on cognitive development, language development and school readiness factor (excluding PTT)**




While the finding that neither programme had an impact on socio-emotional development holds for both the older and younger subsamples, we do find some evidence of age heterogeneity in programme impacts on anthropometric measures (z-scores of weight for age, height for age, BMI for age and weight for height). There is a positive and statistically significant effect of HIM+FE on the three measures linked to weight (weight for age, BMI for age and weight for height) for the older subsample, which for two of these measures (BMI for age and weight for height) do not hold for the younger subsample (although in all cases, the two estimates are not statistically significantly different from one another; Appendix Table E.32). This is an interesting finding, especially in light of our finding that the HIM+FE programme-related improvements in CLS are also concentrated amongst the older subsample.

***Other child characteristics***

In addition to age, we find evidence of heterogeneity in programme effects by baseline levels of stimulation at home and sex. Appendix Table E.29 shows that positive and significant impacts of HIM+FE on CLS and pre-literacy only hold for a subsample of children who had higher levels of stimulation in the home at baseline, as measured by the FCI (see Section 3.2 for a description). This suggests that the effectiveness of preschool-based child development programmes may depend in part on good practices in the home, perhaps due to complementarities in the impacts that the two have on child development.

The home environment also seems to matter when considering impacts of HIM+FE on nutrition. However, here the direction of effects suggests that the nutritional component of the intervention could substitute for a poorer environment in the household: we find larger effects of HIM+FE on measures linked to weight for children who come from

households with lower levels of home stimulation (at baseline) and where the mother has less education. Further, while the effects of HIM+FE on the nutrition measures appear to be concentrated amongst the girls (Appendix Table E.32), we find that the effects on CLS and pre-literacy are only statistically significant for the boys (Appendix Table E.29).

### *5.4.2 Heterogeneity of programme impacts by HI characteristics*

Next, we look at heterogeneity of HIM+FE impacts by HI characteristics. We might expect that smaller or better managed HIs (at baseline) were better able to make use of the resources and training offered through the interventions or, alternatively, that lower quality HIs (at baseline) had more to gain. We consider heterogeneity in programme effects by the size of the HI and an aggregate ICBF quality score (constructed using factor analysis from data on how many of the ICBF criteria the HI met) at baseline. The results show that the improvements in child development outcomes (CLS and pre-literacy) observed for the HIM+FE group were concentrated heavily amongst children in smaller HIs (effect size=0.171, p=0.027) rather than larger ones (effect size=–0.116, p=0.941). The difference between the two estimates is significant (p=0.066) (Appendix Table E.29). The results are again reversed for nutrition outcomes: we observe a significant effect of HIM+FE on weight-based anthropometric measures amongst children in larger HIs only. This could suggest that larger centres were better able to integrate the nutritional improvements into their operations (Appendix Table E.32).

### *5.4.3 Heterogeneity of programme impacts by teacher characteristics*

We find large differences in estimated HIM+FE effects on both child development outcomes measures (CLS and pre-literacy) for pupils of more and less experienced teachers: the effect is large, positive and statistically significant for pupils of more experienced teachers (effect size of 0.323 of a standard deviation, p=0.001 for CLS), but not for less experienced teachers (Appendix Table E.31). Further, the difference between the estimates for these two subsamples is highly statistically significant (p=0.013). This may indicate that more experienced teachers were able to make better use of the training and extra resources that the intervention provided, which resulted in bigger impacts on the children in their classes.

Heterogeneity in HIM+FE effects on nutrition by teacher characteristics is in the opposite direction from that in effects on measures of child development. We find some evidence that HIM+FE has larger and significant effects on the three weight-based indicators (weight for age, BMI for age and weight for height) of children with less experienced teachers (as well as more educated ones; together these suggest younger teachers) (Appendix Table E.34). This is similar to what we observed for heterogeneity in HIM+FE effects on anthropometric outcomes and could again suggest that the nutrition component of HIM+FE can substitute for otherwise less favourable environments.

### *5.4.4 Heterogeneity of programme impacts by compliance to the planned intervention*

As discussed in Section 4, not all HIM fully implemented the programme; most did not hire the intended number of new FTE professionals in the implementation of HIM and there was variation in the numbers and proportions of teachers who attended the FE pedagogic training. It is, therefore, possible that the effects we detect underestimate the potential of the programmes. In order to explore this, we check whether programme effects differ across more and less compliant centres. Since compliance with the

intervention is endogenous, however (i.e. it is likely correlated with other factors that affect the outcomes of a centre), these estimates cannot be interpreted as causal, but only as an indication of whether there is a correlation between child outcomes and programme compliance.

For the HIM intervention, we define more compliant centres as those that achieved at least half of the planned FTE professional-to-child ratio for all three categories of professional (i.e. they had at least one FTE socio-emotional expert and one nutritional expert per 400 children and one FTE teaching assistant per 100 children). This is the case for 34 of the 80 HIM centres. More compliant FE centres are defined as those in which a third or more teaching staff participated in the FE pedagogic training. This is the case in 25 of the 40 HIM+FE centres.[16] We compare outcomes for more and less compliant centres relative to the control group using a regression framework. Results are reported in Table 9. Overall, while we find no differences in impact of the HIM programme between more and less compliant HIM centres, there is some indication that HIM+FE effects are concentrated in centres that showed more compliance with the HIM hiring requirements. The differences are non-zero but very imprecisely estimated and not statistically significant. We do not find differences by FE compliance in the HIM+FE centres, but (as indicated in the footnote) the FE compliance measure should be treated with caution.

**Table 9: Heterogeneity in programme impacts on child development by programme compliance**

| | HIM compliance, HIM centres | | HIM compliance, HIM+FE centres | | FE compliance, HIM+FE centres | |
|---|---|---|---|---|---|---|
| Child development factor | Difference | p-value | Difference | p-value | Difference | p-value |
| Cog, Lang and Sch (all measures, limited sample) | –0.003 | 0.982 | 0.0869 | 0.438 | 0.024 | 0.841 |
| Cog, Lang and Sch (exc. PTT, full sample) | –0.017 | 0.865 | 0.146 | 0.142 | –0.009 | 0.929 |
| Cog, Lang and Sch (exc. PTT, full sample) | 0.018 | 0.870 | 0.094 | 0.405 | 0.024 | 0.839 |
| Pre-literacy skills | 0.050 | 0.602 | 0.139 | 0.160 | –0.002 | 0.988 |

Notes: Table 9 shows the adjusted difference between the more compliant and less compliant centres (programme impact in more compliant centres relative to control – programme impact in less compliant centres relative to control). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$: significance level at which we can reject the hypothesis that this adjusted difference is zero. All p-values adjusted for clustering at the centre level. See notes to Table 4 for description of specifications and outcome measures. More compliant HIM centres are defined as those that achieved at least half of the planned FTE professional-to-child ratio for all three categories of professional (34 centres). For FE, we define more compliant centres as being those who a third or more of their teaching staff participated in the FE pedagogic training (25 centres).

[16] For the purposes of examining heterogeneity in impacts on child development by compliance, we focus on the components of the FE programme that are hypothesised to have an impact on these outcomes (the reading and pedagogic components). Further, we are not able to construct a centre-specific compliance measure, as we do not have disaggregated implementation data for this component. Our measure of FE compliance therefore reflects compliance with the pedagogic training component. Even this measure should be treated with caution, as with the data that we currently have, we are only able to estimate proportion of teachers in training relative to the total number of teachers. A better measure would be relative to the number of eligible teachers within the centre (those teaching children 18–36 months).

## 5.5 Mechanisms

We now examine the impact of HIM and HIM+FE on a core set of intermediate outcomes, which we hypothesise as the mechanisms through which the programmes can improve children's outcomes.

The specific intermediate outcomes we are interested are shown in Figure 2. Outcomes linked to child developmental outcomes include:

1. Improved teaching practices and routines in the class
2. Improved quality of care service
3. Improved routines (reading in particular) in the centre
4. Improved routines (reading in particular) in the house

And linked to child nutrition outcomes:

1. Improved food manipulation and processing in the centres
2. Improved content of servings in the centre
3. Improved food intake by children in the centres
4. Improved nutritional habits in the homes or lower food insecurity

We hypothesise that HIM and HIM+FE would have to improve these intermediate outcomes in order to have substantial positive effects on children's developmental and nutritional outcomes.

### *5.5.1 Teaching practices, routines and overall quality of care in the centres*

We hypothesise that teaching practices, routines and quality of care in HIM classes would improve as the result of recruitment of pedagogic assistants and socio-emotional experts. We would expect these to improve in HIM+FE classes mainly as the result of the pedagogical training component (see Figure 2).

Our most comprehensive measures of these dimensions are sub-scales of the ECERS-R and ITERS-R (Harms et al. 2003), described briefly in Section 3.2. These are measures of the care environment for children: ECERS-R is suitable for classes with children 2 years of age and up, while ITERS-R for classes with under-2s. They capture various dimensions of quality of care relevant here: (1) space and furnishing, (2) personal care routines, (3) language and reasoning, (4) learning activities and (5) interaction and social development. Appendix A shows the complete set of items that make up ECERS-R and ITERS-R sub-scales; they are similar, but individual items have been adjusted to be age-appropriate. Both scales have been shown to have high predictive power for child development in several domains of cognitive (Burchinal et al. 2000) and socio-emotional development (Sylva et al. 2006) and have been extensively used at the international level. Each sub-scale is completed by a fully trained professional who performs a detailed observation of the HI for a sufficient period of time to observe the different routines and activities undertaken throughout the day.

Due to logistical and budgetary constraints, we could only conduct ECERS-R and ITERS-R assessments on a subsample of 216 classrooms in 54 HIs randomly selected, stratifying by city. Almost 60 per cent of these classrooms were administered ECERS-R (because children were older than 2 years), and the remaining 40 per cent were administered ITERS-R (0–2-year-old children in the class). At follow-up, half of the

classrooms we measured were also observed at baseline;[17] the other half corresponded to classrooms attended by children in the sample at follow-up only. We can therefore study not only changes within a classroom over time, but also the extent to which improved processes scores are relevant mediators for children's developmental outcomes.

In order to construct the main measures for the analysis, we applied exploratory factor analysis to total scores for each individual components of each sub-scale (aggregated in accordance with the ECERS-R or ITERS-R methodology). We then standardised the first factor for each sub-scale relative to the baseline values (i.e. using baseline mean and standard deviation).[18]

Appendix Tables E.36 and E.37 show correlations between raw scores on individual sub-scales in baseline and at follow-up. In line with expectations, there is a high prevalence of strong, positive and statistically significant correlations between the sub-scales at both baseline and follow-up, suggesting that classes that perform well in one domain of quality of care are also likely to perform well in other domains.

Table 10 shows estimates of the HIM and HIM+FE programme effects on each of the relevant ECERS-R or ITERS-R sub-scales measured at the class-level, controlling for some key characteristics at baseline, including number of children registered and staff-to-child ratio at the centre, as well as baseline values of the ECERS-R or ITERS-R sub-scales. We include all classes in which ECERS-R or ITERS-R was administered at endline and which are in a centre where ECERS-R or ITERS-R was administered to at least one class at baseline. We combine the ECERS-R and ITERS-R scores to construct the outcome and baseline measures and further use mean scores at the centre level as baseline controls. This yields a sample of 145 classes.

There are statistically significant impacts of the HIM+FE programme across the majority of the sub-scales in Table 10, including personal care, language and reasoning, activities and interactions; the impact on interaction is also there for the HIM programme. The largest and most statistically significant effect is the negative impact of HIM+FE on personal care routines of 83 per cent of a standard deviation.[19] The effects on the other three sub-scales are all positive, significant at the 10 per cent level and very similar in magnitude at about half of a standard deviation. These HIM+FE effects are entirely consistent with the core emphasis in the pedagogical training implemented as part of HIM+FE on focusing efforts on productive learning routines and away from non-productive personal care routines (discussed more in the next section).

[17] Collected for a total 119 classrooms in all 54 centres at follow-up.
[18] We standardise relative to the baseline rather than the control group due to small sample sizes.
[19] Items in the Personal Care Routines sub-scale include greeting and departing; meals and snacks; nap and rest; toileting or diapering; health practices; and safety practices. Items in the Activities sub-scale include fine motor; art; music and movement; blocks; sand and water; dramatic play; nature and science; math and numbers; use of TV, video, or computers; and promoting acceptance of diversity. (See Appendix A.)

**Table 10: Environment rating scales**

| | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | N |
|---|---|---|---|---|---|---|---|
| Space+furnishings | –0.136<br>(0.279) | 0.629 | –0.196<br>(0.374) | 0.603 | –0.060<br>(0.252) | 0.813 | 144 |
| Personal care | –0.074<br>(0.309) | 0.812 | –0.834**<br>(0.375) | 0.031 | –0.760**<br>(0.349) | 0.034 | 138 |
| Language+reasoning | 0.086<br>(0.303) | 0.778 | 0.428*<br>(0.244) | 0.086 | 0.342<br>(0.216) | 0.119 | 145 |
| Activities | 0.148<br>(0.260) | 0.574 | 0.506*<br>(0.254) | 0.052 | 0.358*<br>(0.213) | 0.100 | 134 |
| Interaction | 0.660*<br>(0.338) | 0.056 | 0.481*<br>(0.286) | 0.100 | –0.180<br>(0.289) | 0.537 | 145 |

Notes: All estimates include baseline controls for number of children registered at centre, staff-to-child ratio at baseline and baseline class-level values for complete set of ECERS-R or ITERS-R sub-scales. Each line of the table shows estimates from a separate regression, with the line-specific sub-scale as the outcome measure. Each outcome is measured using the first factor from exploratory factor analysis of scores on all items making up each sub-scale, standardised relative to the baseline. Variation in N across sub-scales is due to missing values for some items in each sub-scale.

Another source of information on classroom routines is the teacher questionnaire. Teachers were asked about their job responsibilities and the details (including frequency) of a large range of activities and routines that they might have performed with their classes during the past week, on how many days and how often per day. We divide these routines into learning activities (which we call 'productive routines'), including telling stories and teaching numbers, and personal care routines ('unproductive routines'), which include brushing hair and pampering children. We see no significant impact of either HIM or HIM+FE on any of these measures (Appendix Table E.38).

In sum, there is evidence of improvements in quality of care in the centres participating in the HIM+FE programme along domains targeted by both of the programmes, but especially the pedagogic training curriculum in HIM+FE. There is a shift of emphasis away from personal care routines towards more productive activities, including different types of play, as the result of the HIM+FE programme.

### *5.5.2 Teacher well-being (satisfaction, depression, burnout)*

Heterogeneity analysis presented in Appendix Table E.31 suggests that effects on child development are particularly significant for more experienced and more educated teachers (as discussed in Section 5.4), but also, contrary to the hypotheses in the pre-analysis plan, ones with poorer baseline well-being indicators (higher burnout and lower job satisfaction). In this section we examine whether the programmes had direct impact on measures of teacher well-being, which include depression, job satisfaction and burn-out. We measured depression using the 10-item version of the Centre for Epidemiological Studies Depression test, job satisfaction using part I of the Early Childhood Job Satisfaction Survey and job burnout using a shorter version of 12 items of the Maslach Burnout Inventory. We constructed total depression, job satisfaction and job burnout scores as indicated by the authors. We find that neither HIM nor HIM+FE had an effect on teacher job satisfaction, depression and burnout. There is a large though imprecisely estimated positive effect of HIM+FE on teacher job satisfaction and some suggestive evidence that HIM might, in fact, have had an adverse effect on teacher depression. What comes out strongly, however, is that relative to HIM, HIM+FE had a significantly more positive effect on teacher job satisfaction and reduction in teacher depression.

**Table 11: Teacher well-being – job-satisfaction, depression, burnout**

| | Adjusted difference HIM-control | p-value | Adjusted difference HIM+FE-control | p-value | Adjusted difference HIM+FE-HIM | p-value | Baseline measure | p-value | N |
|---|---|---|---|---|---|---|---|---|---|
| Teacher job satisfaction | –0.113 | 0.203 | 0.066 | 0.452 | 0.179** | 0.033 | 0.263 | 0.000 | 847 |
| | (0.089) | | (0.088) | | (0.084) | | | | |
| Teacher burnout | –0.012 | 0.895 | –0.097 | 0.287 | –0.085 | 0.334 | 0.270 | 0.000 | 847 |
| | (0.092) | | (0.091) | | (0.088) | | | | |
| Teacher depression | 0.324*** | 0.001 | 0.050 | 0.603 | –0.275*** | 0.003 | 0.118 | 0.005 | 847 |
| | (0.097) | | (0.096) | | (0.093) | | | | |

Notes: Standard errors in parentheses below estimated treatment effects. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$: significance level at which we can reject the hypothesis that (1) the treatment group mean is different from the control group mean (two-tailed test) for comparison of HIM or HIM+FE versus control, or (2) HIM+FE mean is different from HIM mean (two-tailed test). Analysis at the level of the teacher (a total of 847 teachers were interviewed). All estimates control for number of children registered at centre at baseline and teaching staff-to-child ratio at baseline as well as baseline values of covariates (averaged at the centre level). All measures are standardised to have a mean of zero and standard deviation of one in the control group.

### *5.5.3 Improved routines at home*

As part of the HIM intervention, each targeted centre should have received a one-time pedagogical endowment for a value of US$52 per child for toys, books and other materials. Additionally, the FE initiative should have provided funding for books for both parents and centres in the HIM+FE treatment group. In this section, we check the hypothesis that as part of the intervention, the children in the treatment groups have received additional toys and books either directly through the programmes or through changing parents' decisions on purchasing or making such toys, and that it resulted in the change in the allocation of the children's free time in the household.

Overall, we find no evidence that either programme had impacts on children's routines at home. Our primary measures of improved routines at home, defined as providing increased stimulation, are the FCI developed by UNICEF. The FCI measures the quality of the home stimulation environment on five sub-scales: (1) play activities (collected separately for activities with different household members), (2) variety of play materials, (3) sources of play materials, (4) household books and (5) household magazines. The FCI was derived from the much longer and more complex tool to measure the quality of the home stimulation environment – Home Observations for Measurement of the Environment – and was designed to be quicker, cheaper and easier to administer in large survey settings. The tool and these sub-scales have been shown to have good reliability qualities as well as good predictive power over child developmental outcomes (cognitive, language and motor), as measured by the Bayley Scales of Infant Development in Bangladesh (Hamadani et al. 2010).

Estimates of the impact of HIM and HIM+FE programmes show no systematic impact of either intervention on the sub-scales of the FCI. Focusing on reading habits at home, specifically targeted by HIM+FE, we again see no effects of either programme on (1) the number of books available for our study children to read, (2) how long they spent in the previous seven days engaged in activities with books with another household member and (3) how long they spent engaged in activities with books by themselves. However, we do see some evidence of an increase in the number of books borrowed for the child in the HIM+FE group since the start of the programme (January 2013), suggesting that perhaps there are some home reading effects that are difficult to detect using the FCI measure (Appendix Table E.40).

### *5.5.4 Improved food manipulation and maintenance in the centre*

Further, we see no evidence that either HIM or HIM+FE had any impact on food manipulation and maintenance in the centres, based on our summary measures. These measures are constructed through exploratory factor analysis of a series of items asking whether the centre's food preparation and dining facilities met the standards required by ICBF for each of four categories: (1) the quality of preparation and distribution of food, (2) the appearance of the food preparation area, (3) the quality of the storage area and (4) the quality of the dining area. At baseline data collection, we already observed relatively high compliance with the ICBF guidelines for food manipulation and maintenance across all the groups of centres, and it does not appear that the intervention has had any further effect on this aspect in the treatment centres compared to the control (Appendix Table E.41).

### *5.5.5 Improved nutritional habits in the homes – lower food insecurity*

As part of the FE upgrade, the children in the HIM+FE centres were supposed to receive supplementary nutrition that increased calorie intake by 15 per cent, improved protein intake with high biological value and improved iron and zinc intake and vitamins A and B, and is complemented with information about healthy habits and nutritional practices provided to children's parents. To investigate whether this resulted in improved nutrition, as measured by food consumption data reported by parents, Appendix Table E.42 summarises the consumption of different food groups by children in each treatment group. We do not see any pattern emerging from these estimates that would suggest that, as measured by parental report, the frequency of consumption of various food groups changed as a result of either intervention. In Appendix Table E.43, we look at the impact of HIM and HIM+FE on two measures of dietary diversity. We see no significant impact of either programme.

One possible explanation is that the data about child nutrition were collected by interviewing the parents, who might not have been fully aware of any changes in the children's nutritional intake in the centres. Thus, it could still be the case that despite the lack of differences in this intermediate outcome, as we find, the anthropometric indicators for the children attending the treatment centres were positively affected by the intervention compared to the control group.

### *5.5.6 Mediation analysis*

Overall, the most significant finding that emerges from analysis of potential mechanisms is that the HIM+FE programme appears to affect the activities undertaken by HIM teachers in HIM+FE centres – shifting focus away from care routines and towards educational and interactive activities. This is consistent with an improvement in overall practices hypothesises in the pre-analysis plan as a mechanism for the impact of HIM+FE on child outcomes (see pre-analysis plan in Appendix C). The pre-analysis plan further hypothesises that, if sufficiently large, improvements in intermediate outcomes will mediate the programme impacts. We now test this hypothesis by exploring whether and to what extent ECERS-R or ITERS-R performance measured in the follow-up round mediates the impacts of the HIM+FE programme (presented in Table 4).

To do this, we re-estimate the main evaluation model (see Section 5.1) but including follow-up round ECERS-R or ITERS-R scores. As discussed in Section 5.5.1, we do not have ECERS-R or ITERS-R measures for all of the classes and therefore can only conduct this analysis on a subsample of 1,259 children attending centres at follow-up where ECERS-R or ITERS-R was administered in at least one class.[20] The results are presented in Table 12, below. For each of the child development outcomes, we present two sets of estimates of the HIM+FE effect – without and then with the ECERS-R or ITERS-R scores. We then assess the change in the HIM+FE coefficient, testing the significance of the difference between the two coefficients using a cluster bootstrap with 1,000 replications.[21]

[20] We take the centre average ECERS-R or ITERS-R score at follow-up as the measure (clustering standard errors at the centre level).

[21] Note that the mediation analysis is conducted only for the HIM+FE effects on measures of child cognitive development and pre-literacy, since this is where we find strongest program impacts.

We find that the ECERS-R or ITERS-R scores are correlated with measures of child development in a direction that is consistent with their being the mediators for the programme impacts. While personal care routines (which are reduced through HIM+FE) are negatively and (in half of the cases) significantly correlated with all of our measures of child development, the correlations with interaction activities are positive and also statistically significant. Despite this consistency in the direction of correlations, however, these measures do not appear to significantly mediate the programme impact[22]: the size of the HIM+FE effect changes very little with addition of these variables, and the difference between the unmediated and mediated programme impacts is not statistically significant.

These results are inconclusive and suggest that more in-depth analysis would be necessary to determine whether and how class quality might act as a mechanism for the HIM+FE impacts.

[22] Note that while the sizes of the (unmediated) HIM+FE impacts remain the same for this smaller subsample as those presented in Table 4 for the whole sample, they are no longer statistically significant due to loss of precision.

**Table 12: Mediation analysis of class quality**

| | Cog, Lang and Sch (all measures, limited sample) | | Cog, Lang and Sch (exc. PTT, full sample) | | Cog, Lang and Sch (exc. PTT, full sample) | | Pre-literacy skills | |
|---|---|---|---|---|---|---|---|---|
| | Unmediated | Mediated | Unmediated | Mediated | Unmediated | Mediated | Unmediated | Mediated |
| HIM | –0.0823 | –0.0510 | –0.0938 | –0.0593 | –0.106 | –0.0737 | –0.126 | –0.101 |
| | (0.100) | (0.102) | (0.0935) | (0.0958) | (0.102) | (0.104) | (0.0964) | (0.0988) |
| HIM+FE | 0.157 | 0.148 | 0.0535 | 0.0654 | 0.150 | 0.144 | 0.0350 | 0.0451 |
| | (0.101) | (0.101) | (0.0801) | (0.0816) | (0.101) | (0.101) | (0.0802) | (0.0834) |
| Personal care | | –0.0901* | | –0.0633 | | –0.0893* | | –0.0406 |
| | | (0.0497) | | (0.0458) | | (0.0500) | | (0.0444) |
| Language+reasoning | | 0.00876 | | –0.0459 | | –0.00436 | | –0.0389 |
| | | (0.0595) | | (0.0493) | | (0.0577) | | (0.0509) |
| Activities | | 0.0672 | | 0.00664 | | 0.0655 | | 0.00960 |
| | | (0.0651) | | (0.0537) | | (0.0657) | | (0.0538) |
| Interaction | | 0.0939* | | 0.112** | | 0.0965* | | 0.0920** |
| | | (0.0531) | | (0.0442) | | (0.0531) | | (0.0440) |
| P-value for reduction in coefficient on HIM+FE | | 0.379 | | 0.644 | | 0.438 | | 0.619 |
| N | 713 | 713 | 1,259 | 1,259 | 713 | 713 | 1,259 | 1,259 |

Notes: Standard errors in parentheses below estimated treatment effects. * p < 0.1, ** p < 0.05, *** p < 0.01: significance level at which we can reject the hypothesis that (1) the treatment group mean is different from the control group mean (two-tailed test) for comparison of HIM or HIM+FE versus control. Table shows unmediated and mediated effects of HIM and HIM+FE on the main child development measures. Mediated effects are estimated by including class quality controls constructed using ECERS or ITERS data.

# 6. Policy and practice recommendations

Results presented in this report imply some policy and practical recommendations to enhance the impacts of both programmes.

First, the Early Childhood Inter-sectoral Board should continue to promote the evaluation of the main ECD programmes in the country. The board was interested in the preliminary results of the current evaluation and was grateful for having evidence from Colombia, rather than having to rely on international evidence for decision-making. However, it is ICBF that runs the programmes and makes the decisions; it has not received the results with the same enthusiasm. We expect to receive an action plan from ICBF regarding their responses to these results.

FE, on the other hand, has already acted upon the evaluation's results and adjusted its strategic plan for investments in 2016, despite the fact that the estimated impacts are mixed. In the future, FE should continue its collaboration with ICBF, since the investments are substantial and the impacts of the investments can benefit thousands of vulnerable children. In particular, regarding nutrition, the evaluation reports show that the additional nutritional contribution from HIM+FE does not seem to improve nutritional outcomes, despite its additional cost. Nutritional interventions require closer monitoring of implementation. In this case, children might have been overcharged with total calories and the intake of simple carbohydrates and sugars increased as centres received more food.

More generally, because malnutrition was not too high at baseline, perhaps FE's nutritional intervention should have been targeted to a more vulnerable population to begin with. In the current evaluation, the effects we observe do not affect the observed indicators of nutritional status. However, there may be other health benefits of having a more balanced diet that we are not picking up with our outcome measures.

Regarding programme design and implementation, several lessons emerge. First, as described in previous evaluations (e.g. Bernal 2015), it is not enough to increase spending or to bring in more people, toys or books. To enhance the quality of the care provided, we need to go well beyond this. The actual teaching practices and daily experiences of children in the classroom need to improve in order for qualifications to be effective. Staffing centres with new professionals without appropriate training, better articulation with other staff members and job descriptions better aligned with the objective of integrated early education services are less effective than ones where routines are improved directly.

Interventions aimed at improving routines and practices directly in the classroom seem very effective in improving children's development, such as training teachers or implementing reading workshops with teachers, children and parents. There is evidence of improvements in quality of care in the centres participating in the HIM+FE programme along domains targeted by both of the interventions. There is a shift of emphasis away from personal care routines towards more productive activities, including different types of play, as the result of the HIM+FE programme.

Training programmes for early education teachers need to incorporate specific recommendations for ways in which purposeful use of play, art, music and exploration

can actually facilitate the child's learning process; teachers do not seem to fully understand how to put the ideas into practice. The training provided in the HIM+FE intervention is a very hands-on programme in which the weekly assignments aim at generating the materials to be used for specific activities with children. This is a good way of putting ideas into practice and incorporating play, art and exploration into the classroom.

FE's reading programme should be replicated in other centres and other early education services, with particular emphasis on workshops for teachers and parents: These have not only generated improvements in child development, but also have a high approval rate from parents and teachers alike.

When designing policy improvements, we need to bear in mind supply-side considerations. In this case, ICBF wanted to include an interdisciplinary team of professionals and pedagogical assistants for each HI. However, there are very few nutritionists and psychologists, especially in small and medium-sized cities. It was therefore very difficult for centres to comply with ICBF regulations in hiring the full quota of professionals, as reflected in the compliance figures. In addition, professionals complained that the wages they were offered were too low. Future policies should take into account the local labour market conditions. In addition, the full quota of pedagogical assistants to reach the intended ratio of one assistant per 50 children was not reached. If centres had complied with hiring guidelines, the ratio of adult to children in classrooms of children older than 2 years of age would be higher, which could improve the pedagogical experiences. At follow-up, the actual ratio in these classrooms was one teacher per 24 children. Although assistants are extremely helpful with personal care routines in classrooms with babies, the experiences in classrooms of older children did not improve as a result of the presence of the new personnel.

In order for the HIM interdisciplinary team to have substantial impacts on child development, we believe that the programme design should be adjusted. First, a training programme for new professionals hired in centres should be implemented, with an emphasis on articulation with other staff members and aimed at preparing personnel for effective integration of early childhood services. Second, new professional staff could strengthen their role as trainers of teachers in (1) detection of risks (e.g. lags in nutrition, cognitive development or socio-emotional development; maltreatment or abuse; and illness); (2) implementation of action plans in cases of developmental risks; (3) involvement with families; and (4) child assessment and pedagogical planning and implementation in the classroom. Third, new professionals should be trained, and a protocol designed, so that they can implement actual changes and action plans in cases of developmental risks detected. For example, the nutritionist should be able to change the menus for children with nutritional deficits or chronic diseases. Fourth, new professionals should work jointly with teachers and centre directors in improving practices in the classroom and design an on-site supervision and coaching protocol that promotes continuous improvement of teaching practices in the classroom. Full compliance with hiring of paraprofessional pedagogical assistants should be enforced in centres. This would allow a better teacher-to-child ratio, particularly in rooms of older children, and thus facilitate better learning opportunities for children.

# 7. Conclusions

In this report, we focus on evaluation of the impacts of the HIM and HIM+FE programmes on child developmental and nutrition outcomes. We also consider the potential mechanisms for these through looking at how effectively the programmes were delivered and at their impacts on intermediate outcomes.

The results suggest that there were important challenges within both programmes in the implementation of the components of the programmes that were intended to raise children's developmental outcomes. The targets for hiring one socio-emotional expert per 200 children and one pedagogic assistant per 50 children were met in only 30 per cent and 39 per cent of the HIM centres, respectively. According to evidence from the qualitative evaluation, obstacles to reaching these targets included scarce labour supply in some areas, physical or office space restrictions, and working conditions, including wages. It also appears that the specialists hired often spent a substantial amount of time on tasks outside the scope of their intended responsibilities, such as covering for absent teachers and fulfilling administrative duties. Compliance with the pedagogic training programme – a core component of the additional FE improvements – was high, with 87 per cent of teachers who participated completing their course portfolios and receiving the programme certificate. We also find that the courses were attended by other staff, including teaching assistants, directors and other senior staff. Within the reading component of the FE improvement, while all of the book endowments were successfully delivered, there were challenges in getting parents to participate in the reading workshops and finding appropriate space in the HI centres to implement the workshops with the children.

Despite this mixed evidence on implementation success of both programmes, our results suggest that the HIM+FE programme had a positive impact on child development; specifically, children's CLS improved by 0.15 of a standard deviation relative to the control group (children attending regular unimproved HI centres). Analysing the impact of the HIM+FE programme on performance on the specific tests that make up this measure, we find that this effect is driven by improvements in performance on tests measuring fluid reasoning and expressive language. The effects hold most strongly for the older children in the sample, and there is some evidence to suggest that the programme was most effective in raising developmental outcomes for children already being raised in more favourable environments with higher stimulation at home and more experienced teachers, and attending smaller childcare centres. Across most of our measures of cognition, the impact of the HIM+FE improvement is significantly greater than the impact of the HIM improvement. We do not find evidence that either programme had any impact on children's socio-emotional development.

How do we interpret the size of these effects? Where does 0.15 of a standard deviation effect fit in the context of the highly publicised and promoted potential of ECD programmes to have high returns? In considering an appropriate comparator, a few key features need to be taken into account. First, the comparison group in this study consists of children who are also in centre-based care and the difference in care received is to do with quality improvements to the centres. In many studies of ECD programmes, the comparison group consists of children not receiving centre-based care and often living in unstimulating home environments with low-quality, informal childcare arrangements. We

would expect there to be much more potential for improvement in the latter than the former case and, therefore, the effects to be larger as well. Second, many of the often-cited success cases, such as the HighScope Perry Preschool programme (Heckman et al. 2010), were implemented at small scale by highly trained specialists. The evaluation here is of a programme implemented nationally and within the government's centre-based care system, by government-employed professionals in the context of a middle-income country. These features greatly reduce the availability of studies to which impact sizes can be compared.

Perhaps the most directly relevant are two evaluations of ECD programmes conducted by Bernal in Colombia. These were linked to a subsidised, community-based childcare programme for low-income children (a poorer group than in our study) and showed that exposure to the programme improved children's cognitive scores (age 3–6) by 0.15–0.3 of a standard deviation, while a training programme for the care providers within these resulted in improvements in cognitive scores (age under 3) around 0.25 of a standard deviation (Bernal and Fernandez 2013; Bernal 2015). In both of these cases a priori, we would expect impacts to be higher than in our evaluation for a number of reasons: (1) children in both of these studies are from more disadvantaged backgrounds than in our study; (2) in the first evaluation, the counterfactual are children who do not receive any centre-based care; and (3) in the second evaluation, the training programme was carefully designed and implemented by professionals in a highly controlled environment. In the context of comparable studies, therefore, our estimated main effect size of around 0.15 of a standard deviation suggests that the HIM+FE programme may have potential as an effective ECD programme.

Given this finding, it is particularly important to disentangle the mechanisms for the effects – which components of the programme are driving the effects and where there is potential for improvement. Our exploration of potential mechanisms suggests first that the HIM programme was ineffective irrespective of compliance; we do not find any significant impacts even when restricting analysis to centres with higher compliance (at least half of the planned FTE personnel hired). Next, though not fully conclusive, our results suggest that the pedagogic training programme implemented as part of HIM+FE could be an important channel for the HIM+FE effects that we find. We see changes in the activities undertaken by HIM teachers in HIM+FE centres – shifting focus away from less productive care routines and towards more productive educational play and other activities. We find little evidence of effects of either programme on routines at home, however, including stimulation and reading habits, though we do see an increase in the number of books borrowed for the children in the HIM+FE group.

Turning to programme components targeting child nutrition outcomes, the implementation data suggest that there were similar problems with hiring the nutritional expert as with hiring the pedagogical assistant and socio-emotional experts; only one in eight HIM centres met the HIM target of having one FTE nutritionist per 200 children We also see no impact of either programme on key related intermediate outcomes – food manipulation and maintenance in the HI centres, as well as nutritional habits at home.

Analysis of programme impacts on child nutritional outcomes suggests that there were some increases in the weight-based anthropometric measures (weight for age, BMI for age and weight for length z-scores) as the result of the HIM+FE programme. There is

also some evidence of a reduction in proportion of overweight children as the result of the HIM programme. As with impacts on developmental outcomes, the HIM+FE effects on weight are largest for older children. In contrast to the impacts on developmental outcomes, heterogeneity analysis suggests that children in more disadvantaged settings gained more weight as a result of the nutrition components: those from households with lower levels of stimulation at baseline, less educated mothers and attending classes with less experienced teachers.

The findings in this evaluation have a number of key policy implications. Firstly, the evidence clearly suggests that in order to improve quality of childcare provision it is not sufficient to simply increase spending or to bring in more people, toys or books. There needs to be improvement in the actual teaching practices and daily experiences of children in the classroom. With respect to nutrition, since on average malnutrition is not very high at baseline amongst children attending HIs, perhaps nutritional intervention should be targeted to a more vulnerable population to begin with. Finally, when designing policy improvements, we need to bear in mind supply-side considerations. In this case, ICBF wanted to include an interdisciplinary team of professionals and pedagogical assistants for each HI. However, there were very few nutritionists and psychologists, especially in small and medium-sized cities, which was a significant impediment to implementation.

## Online appendixes

Note to the reader: Online appendices are provided as received from the authors. These have not been copy-edited or formatted by 3ie.

**Online appendix A: Survey and implementation instruments – can be accessed here**
http://www.3ieimpact.org/media/filer_public/2017/07/27/appendix-a-columbia-early-years-intervention-ow41162.pdf

**Online appendix B: Attrition analysis, descriptive statistics and balance check for randomization – can be accessed here**

http://www.3ieimpact.org/media/filer_public/2017/07/27/appendix-b-columbia-early-years-intervention-ow41162.pdf

**Online appendix C: Pre-analysis plan – can be accessed here**
http://www.3ieimpact.org/media/filer_public/2017/07/27/appendix-c-columbia-early-years-intervention-ow41162.pdf

**Online appendix D: Econometric methods – can be accessed here**
http://www.3ieimpact.org/media/filer_public/2017/07/27/appendix-d-columbia-early-years-intervention-ow41162.pdf

**Online appendix E: Results tables – can be accessed here**
http://www.3ieimpact.org/media/filer_public/2017/07/27/appendix-e-columbia-early-years-intervention-ow41162.pdf

## References


Bayley N, 1993. *Bayley scales of infant development: Manual*, 2nd edition.

Bernal, R and Camacho, A, 2012. La política de primera infancia en el contexto de la equidad y movilidad social en Colombia. *Documentos CEDE*. Available at: <http://ideas.repec.org/p/col/000089/010313.html>

Bernal, R and Fernández, C, 2013. Subsidized childcare and child development in Colombia: Effects of Hogares Comunitarios de Bienestar as a function of timing and length of exposure. *Social Science & Medicine*, 97, pp.241–249.

Bernal, R and Quintero, C, 2014. Characterization of children younger than five in Colombia based on the Longitudinal Colombia Survey ELCA. Unpublished manuscript.

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

Burchinal, MR, Roberts, JE, Riggins Jr, R, Zeisel, SA, Neebe, E and Bryant, D, 2000. Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development*, 71(2), pp.339–357.

Caldwell, BM and Bradley, RH, 2003. *Home Observation for Measurement of the Environment: Administration Manual*. Tempe, Arizona: Family & Human Dynamics Research Institute, Arizona State University.

Comisión Intersectorial de la Primera Infancia, 2012a. Estrategia de Cero a Siempre: Presentación. 3 May 2015. Available at: <http://www.deceroasiempre.gov.co/QuienesSomos/Documents/Presentacion-estrategia-De-Cero-a-Siempre.pdf>

Comisión Intersectorial de Primera Infancia, 2012b. Prosperidad para la Primera Infancia. Available at: <http://www.deceroasiempre.gov.co/QuienesSomos/Documents/Cartilla-Atencion-Integral-prosperidad-para-primera-infancia.pdf>

Danzer, VA, Gerber, MF, Lyons, TM and Voress, JK, 1991. *DABERON Screening for School Readiness–Second Edition (DABERON-2)*. Pro-Ed (Firm).

Diamond, A and Taylor, C, 1996. Development of an aspect of executive control: development of the abilities to remember what I said and to 'do as I say, not as I do'. *Developmental Psychobiology*, 29(4), pp.315–334. http://doi.org/10.1002/(SICI)1098-2302(199605)29:4<315::AID-DEV2>3.0.CO;2-T

Dunn, LM, Padilla, ER, Lugo, DE and Dunn, LM, 1986. *Test de Vocabulario en Imágenes Peabody (TVIP)*. Circle Pines, Minnesota: AGS Publishing.

Fernald, LCH, Gertler, PJ and Neufeld, LM, 2008. Role of cash in conditional cash transfer programmes for child health, growth, and development: an analysis of Mexico's Oportunidades. *The Lancet*, 371(9615), pp.828–837.

Fernald, LCH, Kariger, P, Engle, P and Raikes, A, 2009. Examining early child development in low-income countries. Washington, DC: The World Bank.

Frongillo, EA, Sywulka, SM and Kariger, PK, 2003. UNICEF psychosocial care indicators project. Final report to UNICEF. Ithaca, New York: Division of Nutritional Sciences, Cornell University.

Hamadani, JD, Tofail, F, Hilaly, A, Huda, SN, Engle, P and Grantham-McGregor, SM, 2010. Use of family care indicators and their relationship with child development in Bangladesh. *Journal of Health, Population, and Nutrition*, 28(1), p.23.

Handal, AJ, Lozoff, B, Breilh, J and Harlow, SD, 2007. Sociodemographic and nutritional correlates of neurobehavioral development: a study of young children in a rural region of Ecuador. *Revista Panamericana de Salud Publica*, 21(5), pp.292–300.

Harms, T, Clifford, RM and Cryer, D, 1998. *Early childhood environment scale-revised edition*. New York: Teachers College Press.

Harms, T, Cryer, D and Clifford, RM, 2003. *Infant/Toddler environment rating scale-revised*.

Heckman, JJ, Moon, SH, Pinto, R, Savelyev, PA and Yavitz, A, 2010. The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94, pp.114–128. doi:10.1016/j.jpubeco.2009.11.001

Heo, KH and Squires, J, 2012. Cultural adaptation of a parent completed social emotional screening instrument for young children: Ages and stages questionnaire-social emotional. *Early Human Development*, 88(3), pp.151–158.

Muñoz-Sandoval, AF, Woodcock, RW, Mcgrew, KS and Mather, N, 2005. *Batería III Woodcock-Muñoz*. Itasca, Illinois: Riverside Publishing.

Peisner-Feinberg, ES, Burchinal, MR, Clifford, RM, Culkin, ML, Howes, C, Kagan, SL and Yazejian, N, 2001. The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*, 72(5), pp.1534–1553.

Romano, JP and Wolf, M, 2005a. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), pp.94–108. http://doi.org/10.1198/016214504000000539

Romano, JP and Wolf, M, 2005b. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), pp.1237–1282. http://doi.org/10.1111/j.1468-0262.2005.00615.x

Schady, N, Behrman, J, Araujo, MC, Azuero, R, Bernal, R, Bravo, D, Lopez-Boo, F, Macours, K, Marshall, D, Paxson, C and Vakis, R, 2015. Wealth gradients in early childhood cognitive development in five Latin American countries. *Journal of Human Resources*, 50(2), pp.446–463. Available at: <http://ideas.repec.org/a/uwp/jhriss/v50y2015i2p446-463.html>

Šidák, Z, 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318), pp.626–633. http://doi.org/10.1080/01621459.1967.10482935

Squires, J, Bricker, D, Heo, K and Twombly, E, 2002a. Identification of social-emotional problems in young children using a parent-completed screening measure. *Early Childhood Research Quarterly*, 16(4), pp.405–419.

Squires, J, Bricker, D and Twombly, E, 2002b. *Ages & Stages Questionnaires: Social-emotional*.

Sylva, K, Siraj-Blatchford, I, Taggart, B, Sammons, P, Melhuish, E, Elliot, K and Totsika, V, 2006. Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, 21(1), pp.76–92.

Walker, SP, Chang, SM, Powell, CA and Grantham-McGregor, SM, 2004. Psychosocial intervention improves the development of term low-birth-weight infants. *The Journal of Nutrition*, 134(6), pp.1417–1423.

WHO Multicentre Growth Reference Study Group, 2006. WHO Child Growth Standards based on length/height, weight and age. *Acta Pædiatrica,* Suppl. 450, pp.76–85. doi:10.1080/08035320500495548

Woodcock, RW, 1977. *Woodcock-Johnson Psycho-Educational Battery. Technical report*. Washington, DC: Distributed by ERIC Clearinghouse.

## Other publications in the 3ie Impact Evaluation Report Series

The following reports are available from http://www.3ieimpact.org/en/publications/3ie-impact-evaluation-reports/3ie-impact-evaluations/

*Can the wounds of war be healed? Experimental evidence on reconciliation in Sierra Leone.* 3ie Impact Evaluation Report 75*.* Cilliers, J, Dube, O and Siddiqi, B, 2018.

*Impact evaluation of the Menabe and Melaky development programme in Madagascar,* 3ie Impact Evaluation Report 74. Ring, H, Morey, M, Kavanagh, E, Kamto, K, McCarthy, N, Brubaker, J and Rakotondrafara, C, 2018.

*Impact evaluation of the Smallholder Dairy Commercialization Programme in Kenya, 3ie Impact Evaluation Report 73.* Bonilla, J, McCarthy, N, Mugatha, S, Rai, N, Coombes, A and Brubaker, J, 2018.

*Impact and adoption of risk-reducing drought-tolerant rice in India*, 3ie Impact Evaluation Report 72. Yamano, T, Dar, MH, Panda, A, Gupta, I, Malabayabas, ML and Kelly, E, 2018.

*Poverty and empowerment impacts of the Bihar Rural Livelihoods Project in India, 3ie Impact Evaluation Report 71.* Hoffmann, V, Rao, V, Datta, U, Sanyal, P, Surendra, V and Majumdar, S 2018.

*How should Tanzania use its natural gas? Citizens' views from a nationwide Deliberative Poll, 3ie Impact Evaluation Report 70.* Birdsall, N, Fishkin, J, Haqqi, F, Kinyondo, A, Moyo, M, Richmond, J and Sandefur, J, 2018.

*Impact evaluation of the conditional cash transfer program for secondary school attendance in Macedonia, 3ie Impact Evaluation Report 69.* Armand, A and Carneiro, P, 2018.

*Age at marriage, women's education, and mother and child outcomes in Bangladesh, 3ie Impact Evaluation Report 68.* Field, E, Glennerster, R, Nazneen, S, Pimkina, S, Sen, I and Buchmann, N, 2018.

*Evaluating agricultural information dissemination in western Kenya*, *3ie Impact Evaluation Report 67.* Fabregas, R, Kremer, M, Robinson, J and Schilbach, F, 2017.

*General equilibrium impact assessment of the Productive Safety Net Program in Ethiopia, 3ie Impact Evaluation Report 66.* Filipski, M, Taylor, JE, Abegaz, GA, Ferede, T, Taffesse, AS and Diao, X, 2017.

*Impact of the Uddeepan programme on child health and nutrition in India, 3ie Impact Evaluation Report 65.* Kochar, A, Sharma, A and Sharma, A, 2017.

*Evaluating oral HIV self-testing to increase HIV testing uptake among truck drivers in Kenya, 3ie Impact Evaluation Report 64.* Kelvin, EA, Mwai, E, Romo, ML, George, G, Govender, K, Mantell, JE, Strauss, M, Nyaga, EN and Odhiambo, JO, 2017.

*Integration of EPI and paediatric HIV services for improved ART initiation in Zimbabwe*, *3ie Impact Evaluation Report 63.* Prescott, M, Boeke, C, Gotora, T, Mafaune, HW, Motsi, W, Graves, J, Mangwiro, A and McCarthy, E, 2017.

*Increasing male partner HIV testing using self-test kits in Kenya, 3ie Impact Evaluation Report 62.* Gichangi, A, Korte, JE, Wambua, J, Vrana, C and Stevens, D, 2017.

*Evaluating the impact of community health worker integration into prevention of mother-to-child transmission of HIV services in Tanzania, 3ie Impact Evaluation Report 61.* Nance, N, McCoy, S, Ngilangwa, D, Masanja, J, Njau, P and Noronha, R, 2017.

*Using HIV self-testing to promote male partner and couples testing in Kenya, 3ie Impact Evaluation Report 60.* Thirumurthy, H, Omanga, E, Obonyo, B, Masters, S and Agot, K, 2017.

*Increasing male partner HIV self-testing at antenatal care clinics in Kenya, 3ie Impact Evaluation Report 59.* Gichangi, A, Korte, JE, Wambua, J, Vrana, C and Stevens, D, 2017.

*Impact of free availability of public childcare on labour supply and child development in Brazil*, *3ie Impact Evaluation Report 58*. Attanasio, O, Paes de Barros, R, Carneiro, P, Evans, D, Lima, L, Olinto, P and Schady, N, 2017.

*Estimating the effects of a low-cost early stimulation and parenting education programme in Mexico, 3ie Impact Evaluation Report 57.* Cardenas, S, Evans, D and Holland, P, 2017.

*The Better Obstetrics in Rural Nigeria study: an impact evaluation of the Nigerian Midwives Service Scheme, 3ie Impact Evaluation Report 56.* Okeke, E, Glick, P, Abubakar, IS, Chari, AV, Pitchforth, E, Exley, J, Bashir, U, Setodji, C, Gu, K and Onwujekwe, O, 2017.

*The Productive Safety Net Programme in Ethiopia: impacts on children's schooling, labour and nutritional status*, *3ie Impact Evaluation Report 55.* Berhane, G, Hoddinott, J, Kumar, N and Margolies, A, 2016.

*The impact of youth skills training on the financial behaviour, employability and educational choice in Morocco, 3ie Impact Evaluation Report 54.* Bausch, J, Dyer, P, Gardiner, D, Kluve, J and Mizrokhi, E, 2016.

*Using advertisements to create demand for voluntary medical male circumcision in South Africa, 3ie Impact Evaluation Report 53.* Frade, S, Friedman, W, Rech, D and Wilson, N, 2016.

*The use of peer referral incentives to increase demand for voluntary medical male circumcision in Zambia, 3ie Impact Evaluation Report 52.* Zanolini, A, Bolton, C, Lyabola, LL, Phiri, G, Samona, A, Kaonga, A and Harsha Thirumurthy, H, 2016.

*Using smartphone raffles to increase demand for voluntary medical male circumcision in Tanzania, 3ie Impact Evaluation Report 51.* Mahler, H and Bazant, E, 2016.

*Voluntary medical male circumcision uptake through soccer in Zimbabwe, 3ie Impact Evaluation Report 50.* DeCelles, J, Kaufman, Z, Bhauti, K, Hershow, R, Weiss, H, Chaibva, C, Moyo, N, Braunschweig, E, Mantula, F, Hatzold, K and Ross, D, 2016.

*Measuring the impact of SMS-based interventions on uptake of voluntary medical male circumcision in Zambia, 3ie Impact Evaluation Report 49.* Leiby, K, Connor, A, Tsague, L, Sapele, C, Koanga, A, Kakaire, J and Wang, P, 2016.

*Assessing the impact of delivering messages through intimate partners to create demand for voluntary medical male circumcision in Uganda, 3ie Impact Evaluation Report 48*. Semeere, AS, Bbaale, DS, Castelnuovo, B, Kiragga, A, Kigozi, J, Muganzi, A, Kambugu, A and Coutinho, AG, 2016.

*Optimising the use of economic interventions to increase demand for voluntary medical male circumcision in Kenya, 3ie Impact Evaluation Report 47.* Thirumurthy, H, Omanga, E, Rao, SO, Murray, K, Masters, S and Agot, K, 2016.

*The impact of earned and windfall cash transfers on livelihoods and conservation in Sierra Leone, 3ie Impact Evaluation Report 46.* Bulte, E, Conteh, B, Kontoleon, A, List, J, Mokuwa, E, Richards, P, Turley, T and Voors, M, 2016.

*Property tax experiment in Pakistan: Incentivising tax collection and improving performance, 3ie Impact Evaluation Report 45*. Khan, A, Khwaja, A and Olken, B, 2016.

*Impact of mobile message reminders on tuberculosis treatment outcomes in Pakistan, 3ie Impact Evaluation Report 44.* Mohammed, S, Glennerster, R and Khan, A, 2016.

*Making networks work for policy: Evidence from agricultural technology adoption in Malawi, 3ie Impact Evaluation Report 43.* Beaman, L, BenYishay, A, Fatch, P, Magruder, J and Mobarak, AM, 2016.

*Estimating the impact and cost-effectiveness of expanding access to secondary education in Ghana, 3ie Impact Evaluation Report 42*. Dupas, P, Duflo, E and Kremer, M, 2016.

*Evaluating the effectiveness of computers as tutors in China, 3ie Impact Evaluation Report 41*. Mo, D, Bai, Y, Boswell, M and Rozelle, S, 2016.

*Micro entrepreneurship support programme in Chile, 3ie Impact Evaluation Report 40.* Martínez, CA, Puentes, EE and Ruiz-Tagle, JV, 2016.

*Thirty-five years later: evaluating the impacts of a child health and family planning programme in Bangladesh, 3ie Impact Evaluation Report 39.* Barham, T, Kuhn, R, Menken, J and Razzaque, A, 2016.

*Effectiveness of a rural sanitation programme on diarrhoea, soil-transmitted helminth infection and malnutrition in India, 3ie Impact Evaluation Report 38.* Clasen, T, Boisson, S, Routray, P, Torondel, B, Bell, M, Cumming, O, Ensink, J, Freeman, M and Jenkins, M, 2016.

*Evaluating the impact of vocational education vouchers on out-of-school youth in Kenya, 3ie Impact Evaluation Report 37.* Hicks, JH, Kremer, M, Mbiti, I and Miguel, E, 2016.

*Removing barriers to higher education in Chile: evaluation of peer effects and scholarships for test preparation, 3ie Impact Evaluation Report 36.* Banerjee, A, Duflo E and Gallego, F, 2016.

*Sustainability of impact: dimensions of decline and persistence in adopting a biofortified crop in Uganda, 3ie Impact Evaluation Report 35.* McNiven, S, Gilligan, DO and Hotz, C 2016.

*A triple win? The impact of Tanzania's Joint Forest Management programme on livelihoods, governance and forests*, *3ie Impact Evaluation Report 34.* Persha, L and Meshack, C, 2016.

*The effect of conditional transfers on intimate partner violence: evidence from Northern Ecuador, 3ie Impact Evaluation Report 33.* Hidrobo, M, Peterman, A and Heise, L, 2016.

*The effect of transfers and preschool on children's cognitive development in Uganda, 3ie Impact Evaluation Report 32*. Gillian, DO and Roy, S, 2016.

*Can egovernance reduce capture of public programmes? Experimental evidence from India's employment guarantee, 3ie Impact Evaluation Report 31.* Banerjee, A, Duflo, E, Imbert, C, Mathew, S and Pande, R, 2015.

*Improving maternal and child health in India: evaluating demand and supply strategies, 3ie Impact Evaluation Report 30*. Mohanan, M, Miller, G, Forgia, GL, Shekhar, S and Singh, K, 2016.

*Smallholder access to weather securities in India: demand and impact on production decisions, 3ie Impact Evaluation Report 28*. Ceballos, F, Manuel, I, Robles, M and Butler, A, 2015.

*What happens once the intervention ends? The medium-term impacts of a cash transfer programme in Malawi, 3ie Impact Evaluation Report 27.* Baird, S, Chirwa, E, McIntosh, C and Özler, B, 2015.

*Validation of hearing screening procedures in Ecuadorian schools, 3ie Impact Evaluation Report 26.* Muñoz, K, White, K, Callow-Heusser, C and Ortiz, E, 2015.

*Assessing the impact of farmer field schools on fertilizer use in China, 3ie Impact Evaluation Report 25.* Burger, N, Fu, M, Gu, K, Jia, X, Kumar, KB and Mingliang, G, 2015.

*The SASA! study: a cluster randomised trial to assess the impact of a violence and HIV prevention programme in Kampala,* Uganda, *3ie Impact Evaluation Report 24.* Watts, C, Devries, K, Kiss, L, Abramsky, T, Kyegombe, N and Michau, L, 2014.

*Enhancing food production and food security through improved inputs: an evaluation of Tanzania's National Agricultural Input Voucher Scheme with a focus on gender impacts, 3ie Impact Evaluation Report 23.* Gine, X, Patel, S, Cuellar-Martinez, C, McCoy, S and Lauren, R, 2015.

*A wide angle view of learning: evaluation of the CCE and LEP programmes in Haryana, 3ie Impact Evaluation Report 22.* Duflo, E, Berry, J, Mukerji, S and Shotland, M, 2015.

*Shelter from the storm: upgrading housing infrastructure in Latin American slums*, *3ie Impact Evaluation Report 21.* Galiani, S, Gertler, P, Cooper, R, Martinez, S, Ross, A and Undurraga, R, 2015.

*Environmental and socioeconomic impacts of Mexico's payments for ecosystem services programme*, *3ie Impact Evaluation Report 20.* Alix-Garcia, J, Aronson, G, Radeloff, V, Ramirez-Reyes, C, Shapiro, E, Sims, K and Yañez-Pagans, P, 2015.

*A randomised evaluation of the effects of an agricultural insurance programme on rural households' behaviour: evidence from China, 3ie Impact Evaluation Report 19.* Cai, J, de Janvry, A and Sadoulet, E, 2014.

*Impact of malaria control and enhanced literacy instruction on educational outcomes among school children in Kenya: a multi-sectoral, prospective, randomised evaluation, 3ie Impact Evaluation Report 18.* Brooker, S and Halliday, K, 2015.

*Assessing long-term impacts of conditional cash transfers on children and young adults in rural Nicaragua*, *3ie Impact Evaluation Report 17.* Barham, T, Macours, K, Maluccio, JA, Regalia, F, Aguilera, V and Moncada, ME, 2014.

*The impact of mother literacy and participation programmes on child learning: evidence from a randomised evaluation in India, 3ie Impact Evaluation Report 16.* Banerji, R, Berry, J and Shortland, M, 2014.

*A youth wage subsidy experiment for South Africa, 3ie Impact Evaluation Report 15.* Levinsohn, J, Rankin, N, Roberts, G and Schöer, V, 2014.

*Providing collateral and improving product market access for smallholder farmers: a randomised evaluation of inventory credit in Sierra Leone, 3ie Impact Evaluation Report 14.* Casaburi, L, Glennerster, R, Suri, T and Kamara, S, 2014.

*Scaling up male circumcision service provision: results from a randomised evaluation in Malawi, 3ie Impact Evaluation Report 13.* Thornton, R, Chinkhumba, J, Godlonton, S and Pierotti, R, 2014.

*Targeting the poor: evidence from a field experiment in Indonesia, 3ie Impact Evaluation Report 12.* Atlas, V, Banerjee, A, Hanna, R, Olken, B, Wai-poi, M and Purnamasari, R, 2014.

*An impact evaluation of information disclosure on elected representatives' performance: evidence from rural and urban India, 3ie Impact Evaluation Report 11.* Banerjee, A, Duflo, E, Imbert, C, Pande, R, Walton, M and Mahapatra, B, 2014.

*Truth-telling by third-party audits and the response of polluting firms: Experimental evidence from India, 3ie Impact Evaluation Report 10*. Duflo, E, Greenstone, M, Pande, R and Ryan, N, 2013.

*No margin, no mission? Evaluating the role of incentives in the distribution of public goods in Zambia*, *3ie Impact Evaluation Report 9.* Ashraf, N, Bandiera, O and Jack, K, 2013.

*Paying for performance in China's battle against anaemia*, *3ie Impact Evaluation Report 8.* Zhang, L, Rozelle, S and Shi, Y, 2013.

*Social and economic impacts of Tuungane: final report on the effects of a community-driven reconstruction programme in the Democratic Republic of Congo*, *3ie Impact Evaluation Report 7*. Humphreys, M, Sanchez de la Sierra, R and van der Windt, P, 2013.

*The impact of daycare on maternal labour supply and child development in Mexico*, *3ie Impact Evaluation Report 6*. Angeles, G, Gadsden, P, Galiani, S, Gertler, P, Herrera, A, Kariger, P and Seira, E, 2014.

*Impact evaluation of the non-contributory social pension programme 70 y más in Mexico*, *3ie Impact Evaluation Report 5*. Rodríguez, A, Espinoza, B, Tamayo, K, Pereda, P, Góngora, V, Tagliaferro, G and Solís, M, 2014.

*Does marginal cost pricing of electricity affect groundwater pumping behaviour of farmers? Evidence from India, 3ie Impact Evaluation Report 4*. Meenakshi, JV, Banerji, A, Mukherji, A and Gupta, A, 2013.

*The GoBifo project evaluation report: Assessing the impacts of community-driven development in Sierra Leone*, *3ie Impact Evaluation Report 3*. Casey, K, Glennerster, R and Miguel, E, 2013.

*A rapid assessment randomised-controlled trial of improved cookstoves in rural Ghana*, *3ie Impact Evaluation Report 2.* Burwen, J and Levine, DI, 2012.

*The promise of preschool in Africa: A randomised impact evaluation of early childhood development in rural Mozambique*, *3ie Impact Evaluation Report 1.* Martinez, S, Naudeau, S and Pereira, V, 2012.

This evaluation assesses the impact of complementary versions of an upgraded early childhood development programme implemented in pre-schools by the Colombian National Family Welfare Agency. It compares the effects of government-implemented improvements to the programme to that of an enhanced version of the programme that included private-sector-funded add-ons for nutritional support for children, pedagogical training for teachers, and a reading component for teachers, children and parents. The authors find that the enhanced version had a positive effect on cognitive and language development, as well as school readiness compared to the version without the add-ons.

**Impact Evaluation Series**

International Initiative for Impact Evaluation
202-203, Rectangle One
D-4, Saket District Centre
New Delhi – 110017
India

3ie@3ieimpact.org
Tel: +91 11 4989 4444

www.3ieimpact.org