**OW4/1162-Evaluation for Centers of Infant Development: an Early Years Intervention in Colombia**

**Pre-Analysis Plan**

Orazio Attanasio, Raquel Bernal, Lina Cardona, Sonya Krutikova, Carlos Medina, Ximena Peña, Marta Rubio-Codina, Marcos Vera-Hernández

1. September 2014
2. **Introduction**

This document outlines a pre-analysis plan—study design, hypothesis to be tested, and data and specifications to be used—for evaluating the impact of the upgrade of *Hogares Infantiles,* implemented by the Instituto Colombiano de Bienestar Familiar (ICBF), and the Fundación Éxito (FE) additions to it. FE is an NGO devoted to the improvement of early childhood services in Colombia, and in particular, nutritional interventions.

Baseline data collection took place between March and June 2013. Follow-up data collection will take place from mid-October to end-November 2014. Since this plan has been written up prior to follow-up data collection, it serves as a pre-commitment for subsequent analysis.

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

Appendix I is a table that summarizes the information collected at baseline and follow up. Additional details on power calculations, the evaluation sample, and baseline balance are provided in the baseline report (Uniandes et al 2014).

1. **Overview of the Study: Interventions and Evaluation Design**

***Hogares Infantiles* (HIs)** are medium-sized childcare centers that provide partly subsidized day care and 60% of daily nutritional requirement to children 2-5 years from low socio-economic backgrounds. They employ between 3 and 6 teachers with some training in early education, each caring for about 30 children. Some centers also enroll children below 2 years. They are run by the national agency for family welfare, the ICBF, and are community-based, in the sense that they are typically located in fairly well-equipped community centers and are run by the parents’ association.

Our study offers a rigorous evaluation, by Randomized Controlled Trial (RCT), of the short-term impacts of two different but complimentary upgrades to HIs on child development outcomes.

*The Interventions*

The **first upgrade** (**HIM**, henceforth) consists of 2 components: (i) hiring a team of professionals, which includes one health/nutrition professional per every 200 children, one expert in socio-emotional development per every 200 children, and one pedagogical assistant per every 50 children to aid with activities in the classroom. The cost of this component is approximately USD 20 child/month, which increases the monthly cost per child from USD 90 to USD 111; and (ii) the delivery of a one-time pedagogical endowment for a value of USD 52 per child for toys, books and other materials. This innovation is fully promoted and financed by the ICBF and is an integral element of the Colombian Government’s National Early Childhood Strategy called *De Cero a Siempre* (From zero to forever), launched in 2011. This innovation constitutes treatment 1, this is to say: **T1 = HIM**.

The **second upgrade** (**FE**, henceforth) **to be added to HIM** consists of 3 components: (i) a nutritional improvement (increased calorie intake by 15%, increased intake of protein with high biological value, and increased intake of iron, zinc, vitamins A and B), complemented with information to parents on healthy habits and nutritional practices, costing USD 11 child/month; (ii) a training program for HI teachers, consisting of 17 3-hour sessions planned, coordinated, and offered by professionals from the Colombian National University, aimed at improving care providers skills to develop pedagogical activities designed to promote infant development, and costing USD 11 child/month; and (iii) a reading program that provides books to centers and parents, and training sessions to parents and teachers on why, what and how to read to children. This upgrade is entirely promoted and financed by FE. Its total cost is USD 22 child/month. Treatment 2 is composed of FE added to HIM, this is to say: **T2 = HIM + FE**.

The Colombian government has committed to the evaluation of the upgrade of HIs since its inception, and hence has allowed us to design a large-scale RCT around the expansion of these enhancement interventions. Members of the research team have been in close communication with the ICBF and other instances in the Colombia government responsible for ECD throughout the research study, to ensure a rigorous evaluation.

*Sample and Evaluation Design*

The evaluation sample consists of 1,987 children 12-36 months in 120 HIs located in seven cities in Colombia: Bogotá, Cali, Medellín, Barranquilla, Bello, Palmira and Itagüi. These cities were chosen because they are amongst the largest in Colombia and hence have the largest number of HI, as well as a population which reflects the geographic, cultural and racial diversity of Colombia, hence adding a certain level of country-representativeness to the sample.

From a universe of 670 HIs in these cities we randomly selected 198, which we organized in groups of 3 by geographical proximity (optimizing distance between the centers to ensure both similar socio-economic conditions and minimal contamination). We selected 40 groups of 3-HIs for inclusion in the study, based on information on attendance by age range. In particular, we needed to guarantee that we had at least 15 children ages 12-36 months (our target age range) to identify impacts. See baseline report for more details on power calculations.

**Then, *within each group of 3*, 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)**, for a total of 40 HIs in each evaluation group. The randomization was performed in November-December 2012.

Baseline data collection on the children, their households and the centers they attend (including surveys to center staff) was collected between March and June 2013, with the exception of 6 centers in Barranquilla, for which data could not be collected until August 2013 (because of the resistance from the teacher’s union to participate in the study). See baseline report for more details.

Follow-up data, initially planned 12 months after baseline, is now scheduled to take place after 15-19 months, during October and November 2014. This decision, motivated by a series of implementation delays (see progress report), aimed to ensure the upgrades had had enough time to consolidate before the evaluation of impacts. The decision followed a number of conversations between FE, the research team and ICBF (the implementing agency).

In particular, the introduction of the interventions experienced the following delays:

* While the FE upgrade was introduced immediately after baseline data collection was finalized in June 2013, its HI teachers’ training component was not completed until May 2014.
* Similarly, while the HIM upgrade started to be rolled-out by ICBF in February 2013, HIs required a substantial amount of time to hire their complete inter-disciplinary team of professionals (nutritionists/health professionals, expert in socio-emotional development, and pedagogical assistants). Follow-up data will inform on the exact date by which HIM was fully rolled-out in each center.

Baseline data collection took place in the HI, except in those instances in which it was not possible to interview the mother of the child in the center, in which case the interview took place in the homes of the respondents. Follow-up data collection will follow the same protocols.

The analysis of baseline data, detailed in the baseline report, shows that the sample is balanced across the evaluation groups. Whilst there are some significant differences observed in children’s nutritional status by group and a few socio-demographic characteristics (e.g. presence of the father, maternal labor supply, enrollment in monthly nutritional controls), none of these differences systematically occur in one of the treatment groups nor point towards a specific (bias) direction.

1. **Hypotheses to be Tested and Data**

We have collected (at baseline) and are planning to collect (at follow-up) a relatively rich set of data which will be used to test a number of hypotheses with regards to the impacts of HIM and HIM+FE and possible mechanisms behind impacts. The hypotheses are presented in groups, which are summarized in Table 1.

**Table 1: Hypotheses Groups**

|  |  |
| --- | --- |
| **A** | **Impact on Children´s Outcomes:** HIM and HIM+FE may have positive average impacts on outcomes for children attending the centers. We group these outcomes into two areas: children´s development and children´s nutritional status.  |
| **B** | **Mechanisms or Intermediate Outcomes:** HIM and HIM+FE are more likely to improve children´s outcomes if the inputs the upgrades aim to introduce in the centers are effectively delivered—i.e. if the inter-disciplinary group of professionals hired at the HI effectively supports nutritional, pedagogical and socio emotional activities at the center; if nutrition (dietary intakes) is improved in the centers; if material resources are received as scheduled; if the reading component is delivered promptly and effectively; if center staff receive the training they are scheduled to receive; if this training teaches center staff skills relevant for their job; if as a result, the quality (and process quality in particular) of the service provided in the centers increases, as well as parenting practices in the home improve (reading, for example).A related relevant intermediate outcome is attendance to the child care center and turnover. HIM and HIM+FE may increase demand in these centers and reduce turnover rates, hence altering the composition of children who attend.  |
| **C** | **Heterogeneity of Impacts:** The individual characteristics of children (age, gender, initial developmental levels, maternal education, baseline level of stimulation in the home), their teachers (education and experience, baseline teaching practices/routines, emotional well-being and job satisfaction), and the centers (size of the child care center and child-teacher/personnel ratio, ICBF’s and international quality standards at baseline) may determine the extent to which children benefit from HIM and HIM+FE. |

**Hypotheses Group A: Impact on Children´s Outcomes**

*HIM and HIM+FE may have positive average impacts on outcomes for children attending the centers.* We group these outcomes into two areas: children’s development and children’s nutritional status.

1. *Children’s development outcomes*

We consider a number of domains within development—namely, some aspects of cognition, language, school readiness, pre-literacy skills, socio-emotional development, and some aspects of executive functioning—and we will collect measures of these. We next list the specific hypothesis on each of them by domain of development, and detail the specific tests (and scales) we will use to measure them and how we will process the data. We would however like to flag three considerations before proceeding.

First, we would like to clarify that we plan to use **factor analysis** (on standardized scores) to determine the most appropriate way of combining the various tests and scales collected in “constructs”. The reason for this is that child development is composed of many different dimensions that are interrelated. For example, even if a vocabulary test like the Peabody Picture Vocabulary Test (PPVT) should be viewed as an achievement test as opposed to a measure of raw ability (given that it measures acquired vocabulary) it nevertheless correlates well with ability (or cognition, the average correlation with standard IQ tests being in the range of 0.62-0.72; Dunn and Dunn 1981). Hence, it is very difficult to establish a priori the most sensible way to organize the data, and we plan to rely on factor analysis to combine data that captures common underlying constructs (data that would be thought to go together on theoretical grounds). To do this, we will follow standard protocols in the use of factor analysis. First, we will construct as many factors (“constructs”) as there are with eigen values larger than 1. Next, we will only use outcomes (scales or tests) with factor loads larger than 0.4 in the construction of these factors.

Hence, the following categorization of scales and tests in domains is for the purpose of illustrating our hypothesis and may be modified as a result of the outcomes of factor analysis. For example, factor analysis may suggest that the tests we are collecting to measure pre-literacy skills and school readiness refer to a common underlying “construct” and should therefore be combined. This “construct” may also include the scales we are considering as a measure of cognition. If so, we will combine them.

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

Second, we would like to note that members of the research team have prior experience in the use in Colombia of the scales to measure child development that we list next and most of them have been used in other ECD studies in Colombia (Bernal et al 2014; Bernal and Fernandez 2013). The final selection of the scales has been made in consultation with a child development psychologist.

Finally, note that the following child development outcomes, listed under Hypotheses Group A, will be collected at follow-up only (as they were not suitable for administration to children at baseline given their age) by direct administration to children by a trained psychologist at the child care center. The psychologists administering the assessments have been hired and trained by a psychologist hired by the research team. The only exception to this is the ASQ:SE, which is collected by maternal report (direct interview with the mother, and was included as part of the household questionnaire) and which was also collected at baseline, since it is appropriate for the entire age range. Children were 18-36 months of age at baseline and will be 36-55 months of age at follow up, depending on the exact time at which they are assessed at follow-up. At baseline we collected different child outcomes, further described under Hypotheses Group C, which will be used as baseline controls, as explained in Section 4.

***Hypothesis A1:*** *HIM and HIM+FE are likely to have a positive average impact on children’s cognitive development, language development, and school readiness, respectively.*

These domains will be assessed using the indicators listed next.

* Cognitive development: some dimensions of cognition will be assessed by scales 5 (concept formation), 12 (fluid reasoning), 17 (memory for words) in the cognitive battery of the **Batería-III Woodcock-Muñoz: Pruebas de Habilidades Cognitivas** (Woodcock and Johnson 1977; Muñoz-Sandoval et al 2005a,c).
	+ The scales will be administered and scored as indicated in the Batería-III Woodcock-Muñoz administration manual. Higher scores indicate higher cognitive abilities. However, given the process of development, scores are also likely to increase with age.
	+ To remove the effect of age, we will standardize the score in 2 different ways: (i) internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods; and (ii) externally, i.e. using CompuScore, the software provided by the test developers. Given that the test has not been normed in Colombia before, we prefer to work with the internally standardized scores. This should be sufficient for the comparison of impacts across children in the control group and in either one of the treatment groups. However, we will construct externally standardized scores for robustness purposes and present them in the appendix. Externally standardized scores are also more likely to allow comparability of scores with those in other studies.
* Language development: we will assess both receptive and expressive language.
* Receptive language will be assessed using the Spanish version of the **Peabody Picture Vocabulary Test (PPVT)**, the **Test de Vocabulario en Imágenes de Peabody (TVIP)** (Dunn et al. 1986). The TVIP measures the extent of vocabulary acquisition of the child by pointing at pictures of objects that represent the stimulus word said by the tester.
	+ The scale will be administered and scored as indicated in the TVIP administration manual. Higher scores indicate higher receptive language development. However, given the process of development, scores are also likely to increase with age.
	+ To remove the effect of age, we will standardize the score in 2 different ways: (i) internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods; and (ii) externally, i.e. using the norms developed from a representative sample of Mexican and Puerto Rico children combined. As noted earlier, we prefer to work with the internally standardized scores (given the unavailability of norms for Colombia) but will replicate the analysis using the externally standardized scores for robustness. Externally standardized scores may facilitate comparability with other studies.
* Expressive language will be assessed using scales 14 (expressive language) in the achievement battery of the **Batería-III Woodcock-Muñoz: Pruebas de Aprovechamiento** (Woodcock and Johnson 1977; Muñoz-Sandoval et al 2005b). The scale requests the child to call out the name of a series of pictures.
	+ The scale will be administered and scored as indicated in the Batería-III Woodcock-Muñoz administration manual. Higher scores indicate higher expressive language development. However, given the process of development, scores are also likely to increase with age.
	+ To remove the effect of age, we will standardize the score in 2 different ways: (i) internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods; and (ii) externally, i.e. using CompuScore. As noted earlier, we prefer to work with the internally standardized scores (given the unavailability of norms for Colombia) but will replicate the analysis using the externally standardized scores for robustness, which may also facilitate comparability with other studies.
* School readiness: we will assess school readiness using a selection of age-appropriate items in the **Daberon-II Screening for School Readiness** test (Danzer et al 1991), a screening tool developed to sample pre-academic knowledge, including knowledge of body parts, color and number concepts, functional use of prepositions, plurals, ability to follow directions, general knowledge, and visual perception.
	+ The scale will be administered and scored as indicated in the Daberon-II examiner’s manual. Higher scores indicate higher school readiness and are also likely to increase with age.
	+ To remove the effect of age, we will standardize the score in 2 different ways: (i) internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods; and (ii) externally, using the norms provided by the test developers, and developed using a sample of children in the US, to compute “Quotients” (in Daberon-II terminology). As noted earlier, we will primarily work with the internally standardized scores (given the unavailability of norms for Colombia) but will replicate the analysis using the externally standardized scores for robustness.

***Hypothesis A2:*** *HIM and HIM+FE are likely to have a positive average impact on children’s socio-emotional development, respectively.*

* Socio-emotional development: we will assess socio-emotional development using the **Ages and Stages: Socio-Emotional Questionnaire** (ASQ:SE), which screens several socio-emotional areas such as self-regulation, compliance, communication, adaptive behaviors, autonomy, affect, and interaction with people, for children 3-66 months by parental report.
	+ The scale will be administered and scored as indicated in the ASQ:SE manual.
	+ While scores should not be age dependent, we will remove any lingering age effect standardizing the scores internally, i.e. using the distribution empirical mean and standard deviation, estimated using non-parametric regression methods.
	+ We will also compute the proportion of children at risk of socio-emotional difficulties in the sample, by using the cut-off points provided by the test developers. We would like to note, that these cut-off points have been developed using a US-representative population and hence may not be applicable to the sample. We will however construct them for comparability with other studies.

***Hypothesis A3:*** *For Hypothesis A1 and A2, the effect of HIM+FE on outcomes is at least as large as the effect of HI—i.e. there is no detrimental effect from adding the components of the FE upgrade on outcomes.*

We do not expect the components (ii) teacher training and (iii) reading program of the FE upgrade to reduce teachers’ and parental time to children in the class or at home, respectively, in a way that is detrimental for child development.

***Hypothesis A4:*** *HIM+FE is likely to have a positive average impact on children’s pre-literacy skills, given its reading component (component (iii)).*

* Pre-literacy skills: we will assess pre-literacy skills using scale 17 (memory for words) in the cognitive battery of the **Batería-III Woodcock-Muñoz**; scales 21 (rhymes) and 14 (expressive language) in the achievement battery of the same test; and the TVIP (receptive language). We will standardize the scores (internal and external standardization) and aggregate them (using factor analysis) following the protocols indicated above.

In addition, we will collect a measure of inhibitory control (and working memory) using the **Pencil Tapping Task** test (Diamond and Taylor 1996). The test offers a continuous score, which we will standardize internally using non-parametric methods as described above. Inhibitory control is part of executive functioning skills, which is a developmental domain on its own but also correlated to a number of other domains. We will use factor analysis to determine the set of outcomes (cognitive, socio-emotional, etc.) our measure of inhibitory control should be included to (i.e. the factor it “better loads with”). As such, this outcome measure will be a component of one or more of hypotheses A1 to A4 above.

1. *Children’s nutritional status*

***Hypothesis A5:*** *HIM and HIM+FE are likely to have a positive average impact on children’s nutritional status, respectively.*

We expect HIM to improve nutritional status through the hiring of a nutritional expert, who is responsible for checking food processing and manipulation, the nutritional content of center meals, and discuss better nutritional habits with parents in the event of malnutrition or risk of malnutrition, amongst others. Component (i) of the FE’s addition directly targets the improvement of the nutritional content of school meals and provides complementary information to parents on healthy habits and nutritional practices.

***Hypothesis A6:*** *The positive impact of HIM+FE on children’s nutritional status is likely to be larger for HIM+FE than for HIM,* given the nutritional improvement (component (i) above) which includes increased caloric protein and vitamin intakes and information on nutrition.

Once the impacts of HIM and HIM+FE are established, we would like to understand how the upgrades improved or not children’s outcomes **(process and mechanisms)**, including the role of compositional effects related to school attendance and turnover rates; as well as whether the upgrades worked better for certain types of children or in certain types of classrooms and centers **(heterogeneity of impacts).**

**Hypotheses Group B: Mechanisms and Intermediate Outcomes**

*HIM and HIM+FE are more likely to improve children´s outcomes if the inputs the upgrades aim to introduce in the centers are effectively delivered—i.e. if the inter-disciplinary group of professionals hired at the HI effectively supports nutritional, pedagogical and socio emotional activities at the center; if nutrition (dietary intakes) is improved in the centers; if material resources are received as scheduled; if the reading component is delivered promptly and effectively; if center staff receive the training they are scheduled to receive; if this training teaches center staff skills relevant for their job; if as a result, the quality (and process quality in particular) of the service provided in the centers increases, as well as parenting practices in the home improve (reading, for example).*

A related relevant intermediate outcome is attendance to the child care center and turnover. HIM and HIM+FE may increase demand in these centers and reduce turnover rates, hence altering the composition of children who attend.

***Common Hypothesis B*:** *We will test the hypotheses that the HIM and HIM+FE upgrades will have an impact on intermediate outcomes. We also hypothesize that these improvements in intermediate outcomes will correlate with (and contribute to—i.e. mediate) the impacts on children’s outcomes described above (Hypotheses Group A).*

This is to say, we hypothesize that HIM and HIM+FE will have to improve intermediate outcomes in order to improve children’s outcomes. If the improvement of these intermediate outcomes is not sufficiently large, the impacts on children’s development and nutritional status will not be statistically significant.

Table 2 summarizes the **specific set of hypotheses B on mechanisms (intermediate outcomes)**. It is a 2x2 matrix where we classify intermediate outcomes according to 2 dimensions:

* the final (children’s) outcomes they are likely to affect—i.e. intermediate outcomes that are required to affect child development and intermediate outcomes that are required to affect child nutritional status
* the upgrades (interventions) that are likely to affect them—i.e. intermediate outcomes that are likely to be affected by HIM and intermediate outcomes that are likely to be affected by HIM+FE.

**Table2. Hypotheses on Mechanisms: Matrix of Observable Intermediate Outcomes by Final Outcomes they will affect and by Type of Intervention**

|  |  |  |
| --- | --- | --- |
|  | **Child Development Outcomes** | **Child Nutritional Status** |
| **HIM, Component (i):**Hiring Professional Team | i. Pedagogic assistant and socio-emotional expert (psychologist) have been hired, how many, and when. ii. Teaching practices and routines have improved in the class iib. Improved quality of care service | iii. Nutritional expert has been hired, how many, and when.iv. Improved food manipulation and processing in the centerv. Improved content of servings in the centervi. Improved food intakes by children in the centers |
| **HIM, Component (ii):**Pedagogical Equipment | vii. Material received and used in the center and used in the class |  |
| **HIM+FE, Component (i):**Increased Nutritional Intake |  | viii. Improved food manipulation and processing in the centerix. Improved content of servings in the centerx. Improved food intakes by children in the centerxi. Improved nutritional habits in the homes/lower food insecurity  |
| **HIM+FE, Component (ii):**Pedagogical Training to Teachers | xii. Teachers have attended training, how many, and contentxiii. Teaching practices and routines have improved in the classxiiib. Improved quality of care service  |  |
| **HIM+FE, Component (ii):**Reading Program | xiv. Materials (books) from program received and how manyxv. Parents and teachers attended reading workshops and how manyxvi. Improved routines (reading in particular) in the centerxvii. Improved routines (reading in particular) in the house |  |

We next describe the intermediate outcomes listed in Table 2 and how we will construct them.

*Intermediate Outcomes*

*i. and iii. Pedagogic assistant and socio-emotional expert (psychologist) have been hired and how many, and when; Nutritional expert has been hired, how many, and when.*

This information is collected in the Child Care Center Principal Questionnaire (Module 8, question 420). We will construct 3 indicators, each one of them equaling 1 if there is at least one staff member from each professional category listed above—pedagogic assistant, socio-emotional expert and nutritionist—in the center. We can construct a second set of indicators equaling 1 if there are as many professionals as required given ICBF’s mandated children/professional ratios.

We will also collect a questionnaire for each professional profile, where we specifically ask them how long they have been employed at the HI and what trainings have they received while being employed at the HI: nutritionist (Module 17, Section IV), socio-emotional professional (Module 16, Section IV) and pedagogical assistant (Module 11, Section IV). We will add up the number of months they have been on the job.

*ii., iib., xiii. and xiiib. Teaching practices and routines have improved in the class; overall quality of care has improved*

We will measure teacher practices, routines, and quality of care in the HI classrooms through the **Early Childhood Environment Rating Scale** (ECERS, Harms, Clifford and Cryer, 1998) and **Infants and Toddlers Environmental Rating Scale-Revised** (ITERS-R, Harms, Clifford and Cryer, 2003).

ECERS is a measure of the care environment for children over 2 years of age. It measures seven dimensions of the child care environment: (i) space and furniture, (ii) personal care routines, (iii) language and reasoning, (iv) learning activities, (v) interaction and social development, (vi) program structure, and (vii) interaction between personnel and parents. Each subscale is completed by a trained professional that performs a detailed observation of the HI for a long enough period of time that must include the different routines and activities developed throughout the day. The ITERS-R is the analogous version of ECERS, adjusted for children 0-2 years.

As we did in baseline (see baseline report), we will construct scores for each scale, as indicated in the administration manual, and two additional scores: the infrastructure subscale, composed of item (i) in the scale; and the processes subscale, composed of the average of items (ii) to (iv). We expect HIM and HIM+FE to improve the process subscales, as well as individual items (iii) language and reasoning, (iv) learning activities, (v) interaction and social development, and (vi) program structure.

As explained in the baseline report, these measures on quality of the child care service were collected in a subsample of 216 classrooms in 54 HIs randomly selected, stratifying by city. Almost 60% of the classrooms were administered ETERS (because children were older than 2 years of age), and the remaining 40% were administered ITERS-R (0-2 years old children in the class). Given the selection of the sample, not all baseline ETERS and ITERS-R scores were balanced.

At follow-up, we plan to collect quality measures on 216 classrooms as follows: half of these (108) will be classrooms for which we also have the baseline measure, ensuring that the 60% ETERS and 40% ITERS-R proportions are maintained; and the other half will correspond to classrooms attended by children in the sample at follow-up. This will allow us to examine both changes within a classroom over time, and to study the extent to which improved processes scores are relevant mediators for higher children’s developmental outcomes. ECERS and ITERs have shown to have a high predictive power on child development through several domains of cognitive (Burchinal et al. 2000a; Burchinal et al. 2000b) and socio-emotional development (Sylva et al., 2007), and hence should be valid intermediate outcomes (and mediators).

Additionally, we will use the information in Section V. Activities in the Hogar Infantil in the Teacher Questionnaire (Module 11), where we ask the teacher about her job responsibilities and the detail (including frequency) of practices and routines in which she engages children in the class, and her response when faced with specific difficulties. We also collected this information at baseline. We will classify routines using factor analysis, following the protocols detailed above. We will correlate these “constructs” to the processes scores from the ITERS-R and ECERS scales to “validate” the teacher reports.

*vii. Pedagogical materials are received in the center and used in class,* as collected in the Principal and Teacher Questionnaires (Modules 8 and 11, respectively). We will also use the relevant questions on materials used for the various learning activities in the ECERS and ITERS scales.

*xiv. Materials (books) from FE reading program are received and how many*, as collected in section VII in the School Principal Questionnaire (Module 8). We will complement this data with administrative records from FE on books delivered per center.

*xii. Teachers have attended FE training and how many*, as collected in section VII in the School Principal Questionnaire (Module 8). We will complement this data with administrative records from FE on trainings delivered per center.

xv. *Parents and teachers attended reading workshops by FE and how many*, as collected in section VII in the School Principal Questionnaire (Module 8). We will complement this data with administrative records from FE on workshops delivered per center and number of attendees.

*xvi. Improved routines (reading in particular) in the center, given that FE’s reading intervention is addressed to teachers*, as measured by subscales (iii) and (iv) in the ECERS/ITERS-R and in the relevant reading routines questions in Section V in the Teacher Questionnaire (Module 11).

*xvii. Improved routines (reading in particular) in the house, given that FE’s reading intervention is also addressed to parents*, as collected using questions on parenting practices from the Family Care Indicators (FCI) developed by UNICEF (which include reading and looking at picture books) and on the number of picture books and books for children in the home that belong to the target child, also from the FCI, play materials subscale. These questions are in Sections XV and XIV, respectively, in the Child at Home Questionnaire (Module 3).

*iv. and viii. Improved food manipulation and processing in the center*, as collected in questions 201-207 and 215 in the ICBF Guidelines Questionnaire, Section V in the Principal Questionnaire (Module 8) and in Section V in the Nutritionist Questionnaire (Module 19).

*v., vi., ix. and ix. Improved content of servings and improved food intakes by children in the centers.*

We will measure food consumption for children in a subsample of 50 HIs. The methodology consists of a visit on a random date, and interviews to 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 and servings for children older than 2, if different. The analysis of these data is based on the ProPAN methodology (Process for the promotion of healthy eating habits of children), a tool developed by UNICEF and Organizacion Panamericana para la Salud (Panamerican Health Organization) to develop, implement and assess interventions and programs to improve diet and eating habits for infants and toddlers. The ProPAN software returns calorie intake by food groups (protein, iron, calcium, vitamin A, C and Zinc) by meal and by age range.

*xi. Improved nutritional habits in the homes/lower food insecurity,* as collected in Section XVII in the Child at Home Questionnaire (Module 3), which enquires about the frequency of consumption of a number of food items (organized by food type) over the past few days. In addition, we included questions on the promotion of food security in the home in Section VI the Nutritionist Questionnaire (Module 19).

A related relevant intermediate outcome is attendance to the child care center and turnover. *HIM and HIM+FE may increase demand in these centers and reduce turnover rates, hence altering the composition of children who attend.* Parental perceptions of increased quality and/or increased resources in the centers may result in increased demand for the services provided by the centers. This may affect the composition of skills of the children attending the centers, which may in turn have an ambiguous impact on the developmental outcomes of children attending the center depending on whether the average skill of their peers increases or decreases.

While this is both an important and interesting research question we are concerned about the extent to which we will be able to rigorously address it given the limited and poor quality of the attendance records kept at the centers. After calling all centers in the sample (see tracking activities report) we have learnt that attendance records are often kept physically and not digitally, and not up-to-date. We will however collect attendance at follow-up and explore the extent to which we can exploit this mechanism in the analysis.

Note also that we will not be able to investigate spillovers and effects on peers due to compositional effects given that we cannot afford to collect data on enough children that were not in the sample (center) at baseline for this purpose.

In addition, as part of a qualitative evaluation—financed with resources from elsewhere—we are collecting process data on a subsample of HIs that will contribute to inform the quality (and reliability) of the intermediate indicators used, and provide insights on the extent to which the above hypothesized mechanisms occurred in reality. The objective of this qualitative component is to characterize the interventions in the field with a detailed tracking of activities and responsible individuals related to the components of the interventions under study. Based on direct observations, in-depth interviews, focus groups with different actors related to the interventions, and play activities with beneficiary children in HIs, we hope to gain a better understanding of the ways in which the interventions are understood, adopted and used directly by HIs.

With this input we hope to be able to (1) better interpret our quantitative estimates of program impacts, (2) identify possible transmission mechanisms of the effects of interest, and (3) propose specific policy recommendations for program improvement based on the joint analysis of the quantitative and the qualitative results.

**Hypotheses Group C: Heterogeneity of Impacts**

*The individual characteristics of children (age, gender, initial developmental levels, maternal education, baseline level of stimulation in the home), their teachers (education and experience, baseline teaching practices/routines, emotional well-being and job satisfaction), and the center (size of the child care center and child-teacher/personnel ratio, ICBF’s and international quality standards at baseline) may determine the extent to which children benefit from HIM and HIM+FE.*

We will examine treatment heterogeneity according to the following dimensions, which we group in three areas: children’s characteristics, teachers’ characteristics, and child care centers’ characteristics.

* *Children’s characteristics*
* **Age:** children will be around 36-55 months at follow up, the proportion of children that will be younger than 3.5 years being around 20%, and the proportion of children that will be older than 3.5 years around 80%. Given the increase in power, and the fact that most of the tests administered are more suitable (more sensitive) for older than younger children, we expect larger impacts on children 42-55 months of age.
	+ **Gender:** we do not know whether to expect differential impacts of the HIM nor HIM+FE by gender and will investigate this hypothesis.
	+ **Initial Developmental Levels at Baseline:** thedirection of this effect is unclear since children with lower initial developmental levels may find it harder to benefit from the extra attention and activities in the center, but also have more scope for improvement in their developmental scores. Children’s initial developmental levels were assessed at baseline using the third version of the **Ages and Stages Questionnaires (ASQ-3)** in Spanish (Squires and Bricker 2009; Squires et al. 1997) and the list of words in the Spanish versions of the **MacArthur-Bates Communicative Development Inventories (CDI)** (Fenson et al. 2000; Jackson-Maldonado et al. 2003). The ASQ-3 screens for delays in five developmental areas: communication, gross motor, fine motor and problem solving. The CDI measures the level of vocabulary acquired by the child (number of words the child can say). Both tests were collected by maternal report (see baseline report for more details on the scoring of these tests, which followed the protocols in the tests’ manuals). We will construct a score using factor analysis and the protocols described above. For the ASQ-3 it is also possible to construct a score by adding up the score in all five developmental areas. We will not use the proportion of children at risk of developmental delay, defined using the external cut-off points because of the low prevalence rates at baseline (all <10%) and because of concerns regarding the suitability of these external cut-off scores to our sample.
	+ **Maternal Education at Baseline:** thedirection of this effect is also unclear since mothers with lower education levels may, for example, find it harder to fully grasp the reading intervention or may face more constraints to attend the sessions; but at the same time there may be more scope for improvement in their child rearing practices, and hence their children’s developmental levels may benefit more. We will use years of education, as collected at baseline, and split the data in education categories (primary completed or less, secondary education (completed or uncompleted), more than completed secondary).
	+ **Level of Stimulation in the Home at Baseline:** in principle we would expect children in homes with lower levels of stimulation to benefit more from the interventions, in particular component (iii) of HIM+FE, since this component specifically aims to promote reading and hence promote parental child rearing practices, which have been shown to be positively associated with child development. However the level of stimulation in the home is also positively associated to maternal education, for example, and hence the same discussion as above applies. We will use the baseline scores for play activities and play materials (FCI variables described above) to measure the level of stimulation in the home at baseline, and split its distribution in quartiles for the analysis of heterogeneity.

Similarly, malnutrition rates at baseline are below 10%, and hence we will not investigate this dimension of heterogeneity.

* *Teachers’ characteristics*
	+ **Education and experience:** teacher with higher skills (higher education level, having received more specialized training, and with more years of experience) are more likely to benefit more from new training, as well as from more and improved pedagogical activities in the center, which may in turn translate in increased children’s outcomes. We collected education, years of experience, and participation in training programs at baseline. We are complementing this information at follow-up with type of educational center attended, to be able to distinguish teachers’ training and/or schooling in institutions with and without credentials.
	+ **Baseline teaching practices and routines**: thedirection of this effect on child development outcomes is unclear since while teachers engaging in more productive teaching practices and routines are more likely to have a higher interest in promoting child development and may be more effective at it, they may also have less scope for improvement in their teaching practices and/or may also be assigned to classrooms with children showing lower developmental levels (to compensate for them). We will use baseline data collected on the number of productive individual and group learning activities and on the number of unproductive routines, (see baseline report for more details on the construction of these variables) and will aggregate them as described above. We will split the distribution of these baseline scores in quartiles for the analysis of heterogeneity.
	+ **Baseline teacher’s emotional well-being (depression), job satisfaction and job burn-out**: we hypothesize that impacts will be on average larger for those children in classes lead by teachers’ with higher levels of emotional well-being, job satisfaction and, related, lower levels of job burn-out. We measured depression using the 10-item version of the **Center for Epedemiological Studies Depression (CES-D)** test (Radloff 1977), job satisfaction using part I of the **Early Childhood Job Satisfaction Survey** andjobburnout 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 (see baseline report for more details). For the analysis of heterogeneity, we will split the distribution of baseline scores for these variables in quartiles.
* *Child care centers’ characteristics*
	+ **Size:** even if the components of HIM and FE are assigned per child, there may still be differential effects by size of the child care center, of unclear sign. In larger centers, with more children and higher children-teacher or personnel ratio, benefits from the upgrades may be lower because of the increased coordination costs. On the other hand, these centers may have a higher ability to adapt to innovations. Baseline center size will be defined as the total number of children enrolled at baseline; children-teacher ratio will be defined as the total number of children per teacher per class at baseline; and children-staff ratio will be defined as the total number of children per pedagogical staff in the center at baseline.
	+ **ICBF’s quality standards and international quality standards**: thedirection of this effect on child development outcomes is also unclear since it is possible that schools with higher standards are more likely to provide services to children with higher developmental levels (see discussion above), and hence may have lower margin of improvement, even if they have more interest and are more able (and willing) to put more effort in providing a service of quality. We will use baseline data collected on school quality standards (ICBF’s and ITERS/ECERS), constructed as defined above and in the baseline report, as well as the measures on routines carried out in the classroom by the teacher. Note that ITERS/ETERS data is only available for a subsample of HIs, as explained above.

The levels of all variables discussed above for the analysis of heterogeneity were balanced at baseline, except for a minor imbalance in children’s ages.

We provide an overview of the information available at baseline and follow-up in Table AI in the Appendix.

1. **Empirical Strategy**

*Basic Specification*

Given the experimental design described in Section 2, we can identify the impacts of HIM (Treatment 1 or T1) and HIM+FE (Treatment 2 or T2) on final outcomes as:

$Y\_{icsl,1}=β\_{o}+β\_{1}T\_{1sl}+β\_{2}T\_{2sl}+γY\_{icsl,o}+X\_{icsl,o}^{'}δ+μ\_{l}+ε\_{icsl,1}$ (1)

where$ Y\_{icsl,1} $is the outcome of interest for individual *i* in class *c* in child care center *s* in city *l* at follow-up (*t*=1); *T1sl* is a dummy equal to 1 if the child care center *s* in city *l* receives Treatment 1 (HIM); *T2sl* is a dummy equal to 1 if the child care center *s* in city *l* receives Treatment 2 (HIM+FE); and $Y\_{icsl,o}$ is the baseline (*t*=0) level of the outcome of interest (or level of the corresponding aggregate construct in the case that the same measure was not administered at baseline and follow-up) for individual *i* in class *c* in child care center *s* in city *l* at follow-up. For child developmental outcomes we will not have the same outcome at baseline and follow-up since we administer different tests given children’s different ages in each data collection round. For these outcomes, we will control for all existing aggregate scores (constructed using factor analysis, as described above) and including all developmental scores and nutritional scores. The purpose of this approach is to maximize efficiency. *Xicsl,0* is a set of basic child characteristics (age and gender), which are also added to improve efficiency (minimize residual variance) given that the sample is balanced across the evaluation groups in terms of observables at baseline, as shown in the baseline report; and $μ\_{l}$ are city dummies. $ε\_{ics1}$ is the random error term, clustered at the child care center level *s* (the unit of randomization).

We can estimate equation (1) by OLS. $β\_{1}$ is the estimated average impact of HIM on outcome $ Y\_{icsl,1}$; (intent-to-treat estimate) and $β\_{2}$ is the estimated average impact of HIM+FE on the outcome (intent-to-treat estimate). In the event of non-compliance in the introduction of the interventions or substantial delays, and assuming non-compliance is not larger than 40%, we would additionally estimate treatment-on-the-treated estimates by instrumenting actual treatment (compliance) with the result of the random assignment (intention-to-treat or randomized treatment variable). We will also assess the extent to which (non-)compliance is correlated with treatment status.

The outcomes of interest will have been standardized as indicated in Section 3. Hence, the coefficient on age is likely to be insignificant and the age variable could be dropped. Leaving it may contribute to reducing residual variance. As described earlier, the outcomes of interest may also be an individual component of a “construct”, depending on the results of the factor analysis model.

*Intermediate Outcomes (Hypothesis Set B)*

We can also use equation (1) to estimate the impact of HIM and HIM+FE on intermediate outcomes.

When intermediate outcomes refer to the teachers (or other child care center staff member) we will replace the set of basic covariates in X with teacher (or other child care center staff member) basic baseline or time invariant characteristics (age and education years). The analogous reasoning applies to intermediate outcomes at the child care center level, in which case the set of covariates in X will be replaced with basic child care center characteristics at baseline or basic time invariant child care center characteristics (namely, size, child-teacher/pedagogical staff ratio, and baseline quality score (if pertinent given the outcome variable)).

*Estimation of Heterogeneous Treatment Effects*

Heterogeneous treatment effects will be estimated by interacting the treatment status dummies (*T1sl* and *T2sl*) with the variable of interest Z, where Z are the variables described under Hypotheses Group C. We will investigate the existence of non-linearities in heterogeneous effects.

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

We have a relatively rich set of outcome measures with which to explore treatment effects. To deal with multiple hypothesis testing we will employ two approaches.

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

The second approach will consist of estimating each outcome (individual test) independently but adjusting p-values for multiple hypotheses testing using the step-down procedure developed by Romano-Wolf on each set of outcomes that constitute an underlying common factor (as determined by factor analysis).

*Survey attrition*

We have engaged in telephone tracking activities before the start of data collection (which achieved a response rate of 60%) and have formally agreed with the data collection company the terms for tracking migrant households. This is aimed at keeping attrition to a rate of 12% or lower.

We acknowledge that a certain level of attrition is unavoidable. We will check that the sample of non-attriters remains balanced on baseline observables (as is the entire sample). We will also check that attrition is orthogonal to treatment status. In the event of a significant correlation between attrition and treatment status we will estimate the determinants of attrition using a probit regression on observables and if feasible use a Heckman selection correction procedure.

*Procedures for Addressing Missing Data*

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

*Questions with Limited Variation*

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

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

*Treatment of Outliers*

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

**References**

Bernal, R., O. Attanasio, X. Peña, M. Vera (2014) **Effects of the Transition from Family-based to Center-based ECD Services on Children** Univiersidad de los Andes, Unpublished manuscript.

**Bernal, R. and Fernández, C. (2013) Subsidized child care and child development in Colombia: Effects of Hogares Comunitarios de Bienestar as a function of timing and length of exposure. Social Sciences & Medicine 97(2013) 241-249**

Burchinal, M. R., Peisner-Feinberg, E. S., Bryant, D., & Clifford, R. M. (2000a). Children's social and cognitive development and child care quality: Testing for differential associations related to poverty, gender, or ethnicity. Applied Developmental Science, Vol. 4, No. 3.

Burchinal, M. R., Roberts, J. E., Riggins, R., Zeisel, S. A., Neebe, E., & Bryant, D. (2000b). Relating quality of center-based child care to early cognitive and language development longitudinally. Child Development, 71(2), 339-357.

Diamond, A., 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,* 315-334.

Dunnn, Ll., E. Padilla, D. Lugo and L. Dunn. (1986). *Manual del Examinador para el Test de Vocabulario en Imágenes Peabody (Peabody Picture Vocabulary Test), Adaptación Hipanoamericana (Hispanic-American Adaption)*, Pearson Assessments, Minneapolis.

Dunnn, Ll. and L. Dunn. (1981). *Peabody Picture Vocabulary Test-Revised*, Dunn Educational Services, Inc., Diablo, California

Fenson L, Bates E, Dale P, Goodman J, Reznick S, Thal D. (2000). “Measuring variability in early child language: Don’t shoot the messenger.” *Child Development,* 71(2):323:328.

Harms, T., Clifford, R. M., & Cryer, D. (1998). Early Childhood Environment Rating Scale. New York: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2003). Infant and Toddler Environment Rating Scale Revised. New York: Teachers College Press.

Jackson-Maldonado D, Thal D, Marchman V, Newton T, Fenson L, Conboy B. (2003). *El Inventario del Desarrollo de Habilidades Comunicativas*.

Marchman VA, Martine-Sussmann C. (2002). “Concurrent validity of caregiver/parent report measures of language for children who are learning both English and Spanish.” *Journal of Speech Language and Hearing Research*, 2002 Oct, 45(5):983-97.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S. and Mather, N. (2005a). *Batería III Woodcock-Muñoz*. Itasca, IL: Riverside Publishing.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S. and Mather, N. (2005b). *Batería III Woodcock-Muñoz: Pruebas de aprovechamiento*. Itasca, IL: Riverside Publishing.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S. and Mather, N. (2005c). *Batería III Woodcock-Muñoz: Pruebas de habilidades cognitivas*. Itasca, IL: Riverside Publishing.

Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. *Appl Psychol Meas* 1977;1:385-401.

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

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

Sylva, K., Taggart, B., Siraj-Blatchford, I., Totsika, V., Ereky-Stevens, K., Gilden, R., et al. (2007). Curricular quality and day-to-day learning activities in pre-school. International Journal of Early Years Education, 15(1), 49-65.

Uniandes, IFS, Banco de la República “Evaluation of Centers for Infant Development: an early years intervention in Colombia. Baseline Report”, July 2014, mimeo

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

WHO. 1983. “Measuring change in nutritional status. Guidelines for assessing the nutritional impact of supplementary feeding programmes for vulnerable groups”, World Health Organization, Geneva.

**Appendix Table A1**

**List of Questionnaires**



**Tests on the Child**



**Center Assessments**

* ETERS, ITERS, on a subsample
* Compliance with ICBF guidelines