Preschool Quality and Child Development

Alison Andrew*†[1,2], Orazio Attanasio[1,2,5], Raquel Bernal[3], Lina Cardona Sosa[1], Sonya Krutikova[1], and Marta Rubio-Codina[1,4]

[1]Centre for the Evaluation of Development Policies, Institute for Fiscal Studies

[2]Department of Economics, University College London

[3]Department of Economics, Universidad de los Andes

[4]Inter-American Development Bank

[5]NBER, FAIR

Abstract

Global access to preschool has increased dramatically yet preschool quality is often poor. We use a randomized controlled trial to evaluate two approaches to improving the quality of Colombian preschools. We find that the first, which was rolled out nationwide and provides additional resources for materials and new staff, did not benefit children's development and, unintentionally, led teachers to reduce their involvement in classroom activities. The second approach additionally trains teachers to improve their pedagogical methods. We find this addition offset the negative effects on teacher behavior, improved the quality of teaching and raised children's cognition, language and school readiness.

JEL Codes: J13, I10, I20, H43.

Keywords: early childhood development, preschool quality, childcare quality.

*We would like to acknowledge excellent research assistance provided by Diana Martínez and the contributions of Carlos Medina and Marcos Vera-Hernandez to the design of this study and of Ximena Peña to both study design and implementation. Ximena passed away in January 2017 and is dearly missed.

†This research was funded by the International Initiative for Impact Evaluation (3ie) and Fundación Éxito. Prof Attanasio acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 695300-HKADeC-ERC-2015-AdG). Ms Andrew, and Dr Krutikova acknowledge funding from the ESRC Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies, Prof Bernal acknowledges funding from the British Academy Visiting Fellowship VF10124. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing up results. Ethics Committees at Universidad de los Andes and University College London approved the study's protocol in 2013.

1 Introduction

There is growing momentum behind investing in early years education in both lower- and higher-income countries. In the UK, for example, government spending on early years care and education has tripled over the past 17 years and, in 2018, the US Congress approved a doubling of funding for subsidized childcare for low-income families (Belfield, Farquharson, and Sibieta 2018; The Economist 2019). Universal access to quality early childhood care by 2030 is one of the Sustainable Development Goals and globally enrollment in pre-primary education has increased from 29% in 1990 to 49% in 2015.[1] In Colombia, the setting of this paper, enrollment rates have increased from 13% in 1990 to 84% in 2015 and, in its 2011 national early childhood strategy, the government committed 0.3% of GDP to delivering high quality early years education - a three-fold increase from the 0.1% of GDP spent in 2005.[2]

This global policy shift is a big achievement in light of the strong evidence that the early years are a critical period for the process of human capital formation (Black et al. 2017; Britto et al. 2017), and a time during which there is great potential to influence children's development through intervention, such as high-quality preschool (Cunha et al. 2006; Heckman 2006; Engle et al. 2007; Heckman et al. 2010; Heckman et al. 2013; Yoshikawa et al. 2013; Heckman and Mosso 2014; Almond et al. 2018). Recent evidence suggests that it is not enough to increase spending on the early years, though; if the money is spent increasing access to poor quality services, it may deliver few benefits or even have adverse effects (Rosero and Oosterbeek 2011; Engle et al. 2011; Britto et al. 2011; Araujo and Schady 2015; Ichino et al. 2019). The great challenge, therefore, is to ensure that services are of sufficiently high quality.

Much remains unknown about which dimensions of quality are most important and how to improve these dimensions at scale. Easily observable characteristics, which make up what is often categorized as *structural*[3] quality have traditionally been used as proxies for quality, although they appear to be poor stand-alone predictors of child development and pupil achievement (Hanushek 2003; Hanushek and Rivkin 2006; Murnane and Ganimian 2014). There is some evidence, instead, suggesting that pedagogical practices and the quality of teacher-child

[1] Figures from World Bank EdStats' "Gross enrolment ratio, pre-primary, both sexes (%)" series. Available from: `https://data.worldbank.org/data-catalog/ed-stats`. This definition gives the total enrollment in pre-primary education, regardless of age, as a percentage of pre-primary age population. It classifies pre-primary education as "Education designed to support early development in preparation for participation in school and society. Programmes designed for children from age 3 to the start of primary education." (ISCED level 02). It should be noted that the preschools we study, while serving children up until age five or six, also serve children under the age of three however there is a lack of internationally and historically comparable data on enrollment rates for programs aimed at younger children.

[2] The government's 2011 spending commitment was made as part of its *De Cero a Siempre* (From Zero to Forever) early childhood strategy. 2005 expenditure from "Government expenditure on pre-primary education as % of GDP (%)" series.

[3] Structural dimensions of quality relate to infrastructure, health, sanitation, safety, and characteristics of teachers and the group of children. Process dimensions, on the other hand, reflect to children's experience, learning and development at preschool e.g. characteristics (frequency, type, quality) of interaction between children and their caregivers.

interaction (which we will refer to as *process* quality) have much higher explanatory power (Araujo, Carneiro, Cruz-Aguayo, and Schady 2016). However, few studies have examined the causal impact of these factors and, among these, the results are mixed (Özler et al. 2018; Yoshikawa et al. 2015). In the context of such ambiguity, governments and other stakeholders struggle to identify ways to ensure that early years spending supports *high-quality* service provision; work in the UK suggests that indicators currently used by the government to allocate funding to preschools are not correlated with child outcomes (Blanden, Hansen, and Mcnally 2017)[4], and in the US evidence on the link between the structural quality factors which make up the National Quality Standard Checklist[5] and child development is very mixed (Hanushek 2003; NICHD Early Child Care Research Network 2002).

In this study, we compare and contrast the effect of targeting different dimensions of preschool quality on children's development. We analyze how both intended and unintended changes in the quantity and quality of teachers' teaching, as well as changes in parental investments, generate these effects. In particular, we designed a three-armed randomized controlled trial (RCT) around the rollout of a nationwide government program in Colombia to improve the quality of public preschools for disadvantaged children. In one arm, public preschools received additional resources (play materials and resources to hire more care-providers) to improve structural quality. In the other, they additionally received a boost to process quality through the provision of teacher training on how to improve the quality of day-to-day interactions with children. To our knowledge, this is the first study to test rigorously and compare directly two substantively different approaches to improving preschool quality. This comparison allows us to weigh up the relative effectiveness of different approaches for improving children's outcomes. Further, our many sophisticated measures combined with latent-factor measurement models lets us to capture the multi-dimensionality of preschool quality and child development, and thus allows us to provide evidence on the mechanisms through which different types of improvements affect childrens outcomes. Finally, this study is built around a real-life national program being implemented in a developing country context. Therefore, in addition to moving forward our understanding of the link between preschool quality and child development, it informs on realistic early childhood policy options for improving outcomes of disadvantaged populations rather than the potential of early childhood programs that studies of well-resourced, typically small-scale and high-income country based interventions speak to.

[4] The same study shows that where children go to preschool does matter suggesting that it is not the case that preschool does not affect child development; just that the factors treated as indicators of quality by the government are unlikely to be the most relevant ones (Blanden, Hansen, and Mcnally 2017).

[5] The checklist was developed by the National Institute for Early Education Research (Barnett, Hustedt, Friedman, Boyd, and Ainsworth 2003; Barnett, Hustedt, Friedman, Boyd, and Ainsworth 2004) and includes teacher qualifications and in-service training, class size, child-teacher ratio, whether there are screening and referral services as well as whether meals are provided.

As part of the trial, 120 public preschools - called "Hogares Infantiles" (HIs) - were randomized into one of three treatment arms: (1) the government's "*Hogar Infantil Mejorado*" (HIM) quality improvement program, which was being rolled-out nationally and consisted of increasing the availability of play materials and number of care-providers, especially teaching assistants;[6] (2) HIM plus an additional program focused on pedagogical training for teachers (HIM+FE); and (3) pure "control" where the implementation of HIM was delayed.

Our results provide stark new evidence that different approaches to increasing investment in the quality of early education can have very different impacts. We show that, in this context, directing money to stand-alone improvements in structural quality (arm 1) is not only ineffective, but potentially harmful since it appears to cause teachers to substitute their effort and involvement in classroom activities for that of the less-qualified and less-experienced new teaching assistants. In contrast, adding well-designed pedagogical training to such spending (arm 2) more than offsets the negative effects on teacher behavior, raises the quality of the learning environment directly observed in classrooms and improves children's cognition, language and school readiness by around 0.15 of a standard deviation (SD), with the biggest gains of around 0.3SD observed for the most disadvantaged children.

We make several contributions to the literature. First, this is one of only a handful of studies to assess the effect of preschool quality improvement programs in a low- or middle-income country at scale and through existing government infrastructure. Such an understanding is crucial for reconciling evidence showing that well-resourced preschool programs (typically small in scale or in high-income countries) *can* have large and long-lasting positive impacts (Cunha, Heckman, Lochner, and Masterov 2006; Heckman, Moon, Pinto, Savelyev, and Yavitz 2010; Heckman, Pinto, and Savelyev 2013; Felfe and Lalive 2018; Havnes and Mogstad 2011) with evidence that large-scale government-run programs in developing countries often do not (Rosero and Oosterbeek 2011; Rao et al. 2012; Rao et al. 2014; Nores and Barnett 2010). Both the programs we study are scalable. The first, HIM, was rolled out to HIs nationwide. We are thus assessing not only a scalable model but the effects of the program when actually implemented at a national scale. The second, FE, was designed so that it could feasibly be expanded to all HIs. The critical teacher training component of FE was run through the Colombian National University and most of the training was carried out through videoconferencing technology to reduce costs and ensure that teachers could fit the training around their busy schedules.

Second, the study informs on *which* dimensions of quality should be targeted through quality-improvement programs. Recent research suggests that enhancements aimed at structural components of quality alone, such

[6] "*Hogar Infantil Mejorado*" means Improved HI.

as, physical infrastructure, teacher-child ratios and the endowment of toys and materials have little or no impact on child development (Bernal, Attanasio, Peña, and Vera-Hernández 2019), in line with findings on primary and secondary education (Glewwe, Hanushek, Humpage, and Ravina 2011; Hanushek and Rivkin 2012). On the other hand, evaluations of programs that, given a minimum level of structural quality, directly focus on improving process quality through, for example, training to improve teacher–child interaction or the adoption of structured pedagogical curricula, have often been found to have positive impacts on child development (Nores, Bernal, and Barnett 2019; Bernal, Martínez, and Quintero 2015; Attanasio, Baker-Henningham, Bernal, Meghir, Pineda, and Rubio-Codina 2018; Araujo and Schady 2015). However, it is difficult to compare the impact of improving structural quality with the impact of also targeting process quality by contrasting impacts found in different studies set in very different contexts. To the best of our knowledge, this is the first randomized study to directly make this comparison.

Third, we go beyond a simple comparison of the causal impacts of the two different approaches and provide evidence on the mechanisms that generate these impacts - both the intended and unintended behavioral responses. An understanding of these mechanisms is critical, especially for predicting how our findings might generalize to other contexts. We show that providing additional support staff may have the unintended negative impact of reducing the time that teachers spend doing educational activities with preschoolers. In contrast, we show that providing teachers with training about early childhood development and support in developing high-quality classroom activities can reverse this negative impact and additionally increases the quality of directly observed classroom activities. These findings may be relevant beyond preschool. For example, while the Tennessee Project STAR program found that while increasing the intensity of teacher-student interactions through reducing class sizes improved children's performance, the addition of teaching aides did not (Hanushek 1999; Gerber, Finn, Achilles, and Boyd-Zaharias 2001). Our findings suggest that a plausible explanation is that teachers may respond to additional human resources in the classroom by reducing their own involvement but that training about the importance of high-quality teacher-pupil interactions might mitigate this.

A key feature of the study, which facilitates all three of these contributions, is more refined measurement than is common in this literature. We use many and sophisticated measures of child development, of preschool quality and of the quality of the home environment. We then use latent-factor measurement models, which explicitly recognize the existence of measurement error, to extract efficient estimates of multiple dimensions of each of these underlying constructs. To measure child development, for example, we use 15 different measures which cover

cognition, language, school readiness, executive function, pre-literacy and socio-emotional skills. We score the measures using methods that use efficiently the information contained in children's responses to every item and then use a factor model to assess the dimensionality of our measures and predict these underlying dimensions. This type of approach is widely applied in the structural literature on child development (Attanasio, Cattan, Fitzsimons, Meghir, and Rubio-Codina 2015; Cunha, Heckman, and Schennach 2010; Attanasio, Meghir, and Nix 2017; Agostinelli and Wiswall 2016) but much less frequently used in program evaluation despite its substantial advantage of reducing measurement error which, given the standard practice of anchoring effect sizes to the standard deviation of the untreated sample, not only reduces precision but also causes attenuation bias in the anchored treatment effect. Our approach to measurement allows us to assess the impact of the two quality improvement programs on multiple dimensions of child development and to do so with precision. Moreover, capturing multiple dimensions of each of three important margins on which preschool teachers and parents might respond to the programs – teachers' involvement in classroom activities, the quality of that involvement and parents' investments in creating a stimulating environment at home – allows us to assess the mechanisms through which the programs impact child development.

The findings in this study are especially pertinent given that the part of the program which we find to be ineffective, and potentially harmful (HIM), is now implemented nationwide. All HI preschools now receive the extra resources for hiring additional staff and play materials, but without the pedagogical training program which we find to be critical for translating this extra expenditure into positive impacts on child development. Our results, therefore, offer evidence on a feasible and scalable way to significantly enhance the efficacy of the current government quality improvement program. A rough assessment of costs suggests that the pedagogical training program could be scaled at a one-off cost of US$5827 per HI and an ongoing cost, for training new teachers and running less-intensive refresher training, of US$2206 per center per year, or an increase of 1.0% above current expenditures.[7]

The rest of the paper is organized as follows. The next section outlines the interventions we study, our study design and our empirical strategy. Section 3 outcomes of interest and how we measure them. In Section 4, we present our results on both child outcomes and mechanisms. Section 5 concludes and discusses policy implications.

[7] All computations in USD henceforth, use the exchange rate in February 2013 (1,800 COP/USD) when the interventions we study were implemented.

2 Interventions, Study Design and Empirical Strategy

2.1 Setting and Interventions

The preschools we consider in this study are *Hogares Infantiles* (HIs) which provide partially-subsidized preschool for low socio-economic status children aged 5 years and younger.[8] HIs are targetted at children whose parent(s) are working and therefore who are at risk of inadequate childcare. The program is well-established; it is the oldest public center-based childcare program in Colombia and has enrolled an average of 125 thousand children per year during the last decade. There are now 1,008 HIs across the country which are run by the government with input from parent associations. The preschools are typically located in fairly well-equipped community centers and employ between 3 and 10 teachers who have some training in early education; each teacher cares for about 30 children. In 2011 the government of Colombia announced that the HIM upgrade of HIs would be a central component of its US$1.28 million *"De Cero a Siempre"* (or From Zero to Forever) program (Bernal and Camacho 2012; Bernal and Ramírez 2019).

2.1.1 HIM

The Government of Colombia's *Hogares Infantiles Mejorado* (HIM, or improved HI) program provides HIs with resources for (1) hiring classroom assistants, nutrition or health professionals, and professionals in child socio-emotional development; and (2) buying toys, books and other pedagogical materials. The funds for the program were provided on a per child basis: each center received $20 per child per month for the hiring component and a one-off payment of $52 per child for the materials. The government provided some guidance on the level of hiring centers should aim to achieve with this money: one full-time socio-emotional expert for every 200 children, one full-time nutritionist for every 200 children and one full-time classroom assistant for every 50 children. Overall, these improvements entailed a considerable financial commitment on the part of the government: a 30% increase in per child expenditure relative to the business-as-usual unenhanced model (from $1,000 to $1,300 per year).

To understand whether HIs indeed spent the additional resources on hiring we utilize data on number of children in a given HI to compute the total budget allocated to each HI to spend on hiring new staff. We then use personnel data, including data on salaries, collected at baseline and endline to calculate what proportion of the budget allocated for hiring the additional personnel was spent by HIs in this way. This exercise suggests that on

[8] Occasionally, HIs take children as young as 6 months when it is "proven that they do not have a responsible adult to care for them". However, the vast majority of children enrolled in HIs are 18 months or older.

average hiring compliance was high with more than 70% of the money allocated for hiring socio-emotional experts, nutritionists and classroom assistants spent in this way across the two treatment arms.

Although HIs spent most of the money allocated as prescribed, the government guidelines on the professional-child ratios that HIs should aim to reach through this hiring appear to have been overly optimistic given actual market wages. The great majority of HIs (90%) had employed a nutritionist, a socio-emotional expert and at least one classroom assistant by endline. However, less than 10% were able to achieve the recommended full-time equivalent personnel-child ratios.

2.1.2 FE

The teacher training and reading programs were designed by a well-established Colombian NGO, Fundación Éxito (FE), in partnership with the Colombian National University, as an additional enhancement to the government improvements.[9] The FE upgrade began in June 2013 and was completed in May 2014.

The teacher training program was planned, coordinated and implemented by expert professionals from the departments of psychology and education at the Colombian National University. The aim of the program was to train teachers on how to design and implement activities that promote children's development. The curriculum covered technical guidelines for early childhood services; child development from 18 to 36 months of age; nutrition; brain development; cognitive development; early literacy; the use of art, music, photography and body language for child development; mathematical concepts during early childhood; and pedagogical strategies during early childhood.

The training was delivered through three components: (1) instruction through 16 monthly 3-hour long sessions delivered via videoconferencing; (2) video tutoring sessions of three hours per week in which participants worked with their tutors on-line on developing and refining classroom activities; and (3) on-site coaching where instructors carried out one classroom observation of participating teachers to provide specific feedback on their content and pedagogical methodology. The program was offered for free but participating teachers incurred costs of transportation to monthly sessions, required internet access and needed materials for preparation of new activities.

The teacher training component of the FE enhancement was implemented between June 2013 and June 2014.

[9]The FE program also included additional nutritional improvements which aimed to increase calorie provision by 15% over the 60% of daily requirements already provided by HIs. We have documented elsewhere that under-nutrition did not appear to be a particular problem in this population at baseline and, perhaps not surprisingly therefore, we find no evidence of HIM+FE having any impact on any measure of nutritional status (Andrew, Attanasio, Bernal, Cordona, Krutikova, Heredia, Medina, Peña, Rubio-Codina, and Vera-Hernandez 2018).

Center directors nominated 2-3 teachers per treated HI to participate, with some additional teachers from the same centers selected to replace teachers who were not able to attend all of the sessions or who dropped out. Administrative records indicate that 114 teachers in the 40 HIs assigned to HIM+FE started the training. Out of these, 99 teachers (or 87%) were certified as having completed it. Although the training was designed for teachers, in some cases other staff, including classroom assistants, directors or other senior staff, additionally participated as well.

The reading program aimed to provide parents and teachers with strategies to motivate their children to read, and emphasized reading as a way for parents to bond with their children. It had three components: reading and music promotion, encouraging effective bonding in families through reading, and building an appropriate environment for reading in the preschools. More specifically, the program involved providing books and book bags to centers, providing training workshops to parents and teachers focused on reading with children as well as running some reading workshops directly with children. The design and delivery of the program were commissioned to a Colombian NGO devoted to promoting good reading habits among youth.

2.2 Study Design and Sample

We designed a three-armed cluster randomized controlled trial around the national rollout of the Colombian government's preschool quality enhancement program to assess both the impact of HIM and HIM enhanced with a teacher training and reading program (HIM+FE). Randomization was at the level of the HI with 40 HIs randomized into each of the three arms: (i) HIM, where preschools received the government quality improvement program in line with all other, non-study, HIs, (ii) HIM+FE, where preschools received the FE enhancements in addition to the HIM program and (iii) a pure control group where the implementation of HIM was delayed. This design allows us to test several hypotheses which we set out in a pre-analysis plan held at the AEA trial registry (AEARCTR-0001246). First, we are able to study whether the government improvement program had an impact on children attending the upgraded centers relative to those in the "business-as-usual" HIs. Second, we are able to evaluate the full impact of the HIM+FE program relative to "business-as-usual" HIs. And finally, we are able to test whether adding the FE component enhances the effectiveness of the government upgrade.

To select the 120 study HIs, we first obtained GPS co-ordinates for the 248 HIs in eight study cities. In order to increase the likelihood of having a balanced sample, we organized HIs into groups of three geographically close

HIs, from which we selected 40 triplets for inclusion in the study based on the HIs having at least 15 children in our target age range (18 to 36 months at baseline). Then, within each triplet, we randomly assigned one HI to the control group, one HI to the HIM treatment group and one HI to the HIM+FE treatment group. Randomization and sample selection were carried out over November-December 2012.

This procedure yielded a final sample of 120 HIs, with 40 in each of the three groups. On average, the HIs in the sample had 48 children between the ages of 18 and 36 months from whom we drew a baseline sample of 15 to 17 children per HI.[10][11] Baseline data were collected between March and May 2013.[12] The total baseline sample consisted of 1,987 children (663 in HIM centers, 663 in HIM+FE centers and 661 in control group HIs). At endline, 18 months later during October and November 2014, we tried to reach all children in the study sample, regardless of whether they were still attending an HI or not, and regardless of the length of their exposure to the programs. The RCT flow chart (Figure B.1) shows how the final study sample was selected.

2.3 Balance and Attrition

As shown in the flow chart in Figure B.1, the attrition rate was relatively low – of the 1,987 children in the baseline sample, we have complete child development assessments for all but 155 (7.80%) – and was not related to treatment assignment (Table A.1). The sample at endline is largely balanced across key child and household characteristics measured at baseline (Table 1). Exceptions include a significantly higher proportion of male children in the two treatment groups than the control group; and that the children in the HIM+FE group are older than those in the HIM group and the control group. We address the two baseline imbalances by controlling for children's age and gender in our empirical strategy.

The majority of children (72.2%) continued attending the same HI throughout the study period; by endline, 9.2% were enrolled in a different HI (mostly one not in the study sample); 13.1% were enrolled in a different public or private child care service; and 5.5% were not enrolled in any type of child care service. The probability that children remained in the same HI was not impacted by treatment status.

[10] HIs were selected so that there was at least 15 children aged 18-36 months at baseline in each. Where there was 15, 16 or 17 children in the target age range, we included them all. When there were more, we randomly selected 17.

[11] Power calculations assumed program effects of 0.20 of a SD for all pair-wise comparisons on cognitive development. We assumed an intra-cluster correlation coefficient of 0.035 which was similar to what we observed in the baseline data for cognitive development (measured by the Ages & Stages Questionnaire and conditional on observables). Given our assumptions, we calculated that we would require 15 children per cluster (HI) to achieve 80% power at a 5% significance level. To allow for attrition between baseline and follow-up data collection (of about 10%), we assessed 17 children per cluster whenever possible.

[12] Baseline data collection happened after the start of the HIM rollout, which began in February 2013, due to delays in data collection. However, we argue in footnote [27] that this is not driving our results.

Table 1: Baseline sociodemographic characteristics and child outcomes by randomization status for analysis sample

	Control	HIM only	HIM+FE	HIM vs. Control	HIM+FE vs. Control	HIM+FE vs. HIM	N
Male	0.469	0.538	0.521	$[p = 0.004]$	$[p = 0.062]$	$[p = 0.582]$	1819
	(0.499)	(0.499)	(0.500)	$\{p = 0.010\}$	$\{p = 0.075\}$	$\{p = 0.590\}$	
Age (months)	29.71	30.06	28.96	$[p = 0.238]$	$[p = 0.067]$	$[p = 0.005]$	1819
	(4.607)	(4.387)	(4.892)	$\{p = 0.334\}$	$\{p = 0.054\}$	$\{p = 0.012\}$	
HH income (million COP)	1345.4	1292.2	1373.3	$[p = 0.406]$	$[p = 0.744]$	$[p = 0.208]$	1819
	(751.626)	(727.624)	(816.559)	$\{p = 0.441\}$	$\{p = 0.690\}$	$\{p = 0.242\}$	
Mother's education (years)	12.69	12.35	12.67	$[p = 0.081]$	$[p = 0.934]$	$[p = 0.083]$	1802
	(2.722)	(2.608)	(2.628)	$\{p = 0.068\}$	$\{p = 0.918\}$	$\{p = 0.087\}$	
Father's education (years)	12.08	12.02	12.16	$[p = 0.780]$	$[p = 0.741]$	$[p = 0.483]$	1699
	(3.034)	(2.986)	(2.962)	$\{p = 0.802\}$	$\{p = 0.722\}$	$\{p = 0.506\}$	
Household size	3.384	3.434	3.239	$[p = 0.625]$	$[p = 0.159]$	$[p = 0.061]$	1819
	(1.639)	(1.613)	(1.522)	$\{p = 0.634\}$	$\{p = 0.174\}$	$\{p = 0.044\}$	
ASQ Child Development Factor Score	-0.0280	0.0830	-0.0547	$[p = 0.307]$	$[p = 0.849]$	$[p = 0.138]$	1817
	(1.021)	(1.023)	(0.951)	$\{p = 0.290\}$	$\{p = 0.809\}$	$\{p = 0.154\}$	
Language Development (MacArthur Bates CDI)	0.0317	0.0284	-0.0363	$[p = 0.990]$	$[p = 0.430]$	$[p = 0.472]$	1819
	(1.009)	(1.011)	(0.964)	$\{p = 0.968\}$	$\{p = 0.460\}$	$\{p = 0.461\}$	
Socio-Emotional Factor Score (ASQ:SE)	0.0201	0.0788	-0.0974	$[p = 0.443]$	$[p = 0.190]$	$[p = 0.028]$	1819
	(1.040)	(0.912)	(1.036)	$\{p = 0.467\}$	$\{p = 0.196\}$	$\{p = 0.028\}$	

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using bootstrap p-values. Baseline characteristics by treatment status for children included in the analysis sample (all children with non-missing child development assessment data at endline). Single-hypothesis two-sided p-values calculated 2 ways: (i) $[bootstrap]$ and (ii) $\{randomization - t\}$. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. ASQ Child Development Factor Score is the first factor from an exploratory factor analysis of the five subscales of the ASQ: Communication, Gross Motor, Problem Solving, Personal-Social and Fine Motor. Socio-Emotional Factor Score is the first factor from an exploratory factor analysis of the seven subscales of the ASQ:SE: Self Regulation, Compliance, Communication, Adaptive Functioning, Autonomy, Affect and Interaction with People. Child development measures age-standardised and scaled to have zero mean and unit variance across the analysis sample.

2.4 Empirical Strategy

We evaluate impacts using an intention-to-treat approach. Thus, our analysis sample includes all study children regardless of whether they attended the center throughout the intervention period. Given the experimental design we estimate the impact of a child's baseline HI being allocated to HIM ($T_{lm}^{HIM} = 1$) or HIM+FE ($T_{lm}^{HIM+FE} = 1$) on final outcomes through OLS:

$$Y_{clm} = \beta_0 + \beta_1 T_{lm}^{HIM} + \beta_2 T_{lm}^{HIM+FE} + X_{clm}\gamma + \epsilon_{clm} \tag{2.1}$$

Where Y_{clm} is the outcome of interest for child c, in child care center l, in triplet m. X_{clm} is a pre-specified set of control variables added to improve efficiency. ϵ_{clm} is the random error term, which we allow to be clustered at the level of the sampling triplet.

Pre-specified baseline controls comprise the child's gender, a set of city dummies, a set of tester or interviewer dummies, and, if applicable, the baseline levels of the outcome of interest.[13][14] We administered different assessments of child development at baseline and endline – endline measures are outlined in Section 3.1 – to ensure that we used age-appropriate instruments. Baseline child-development controls therefore include five subscales of an extended version of the ASQ-3 to measure Communication, Gross Motor, Problem Solving, Personal-Social and Fine Motor (Squires, Bricker, and Twombly 2009); MacArthur Bates Communicative Development Inventories (Jackson-Maldonado et al. 2013; Jackson-Maldonado et al. 2003) to measure language development;[15] and ASQ:SE (Squires, Bricker, and Twombly 2002) to measure socio-emotional development.

We report β_1, the average impact of HIM relative to control, β_2, the average impact of HIM+FE relative to control, and $\beta_2 - \beta_1$, the average impact of HIM+FE over and above HIM. We construct standard errors using a block bootstrap, re-sampling the 40 randomization triplets with replacement (5000 iterations). We construct two-sided single-hypothesis p-values in two ways: (1) through the same block bootstrap; (2) through randomization inference (randomization-t). Randomization inference has the advantage of providing exact tests where the size is unaffected by sample size, the distribution of errors and the joint distribution of the dependent variables (Young 2018).[16]

[13] Pre-analysis plan available at the AEA RCT Registry (ID AEARCTR-0001246).

[14] The inclusion of tester or interviewer dummies depends on whether the outcome was measured by the tester (psychologist) or the household interviewer.

[15] We extended the ASQ for each age-specific questionnaire by adding the last three non-overlapping items in each subscale from the age-specific questionnaire below and the first three non-overlapping items in each subscale from the age-specific questionnaire above. This was to ensure the instrument had sufficient discrimination over the entire support of baseline child development.

[16] Although, randomization inference only provides exact tests when the null hypothesis in question is sharp, in our case this

Whenever we test the same hypothesis (i.e. the difference between any two treatment arms) on multiple conceptually-similar pre-specified outcomes we also present p-values that are adjusted for multiple testing across these outcomes. To do this we use the stepwise procedure described in List, Shaikh, and Xu (2016) which, building on Romano and Wolf (2005) and Romano and Wolf (2010), provides balanced asymptotic control of the family-wise error rate. In running the procedure we use the block bootstrap described above, studentizing by the bootstrap standard error, to simulate the distribution of studentized test statistics under the assumption that all null hypotheses are true. Importantly, this method accounts for interdependence between hypothesis tests, which increases the tests' power compared to classical methods.

3 Outcomes and Measurement

To understand the impact of two approaches to improving preschool quality on child development and the mechanisms that generated any impacts, we seek to estimate the effects of the two approaches on three conceptual constructs: child development, preschool quality and the quality of the home environment. Each of these constructs is potentially multidimensional. As we discuss below, our data is very rich; we collected a variety of measures related, although not perfectly, to each of these constructs. Therefore, we need a framework that relates these conceptual constructs to the instruments that measure them and which, in turn, allows us to use information contained in our multiple error-ridden measurements to obtain predictions of the underlying, potentially multi-dimensional, constructs. In this section, we describe our methodology which uses an explicit model of measurement and measurement error. The explicit consideration of a measurement system allows us to use all the available information efficiently thereby reducing measurement error. Measurement error in outcome variables is of particular concern when the size of treatment effects is anchored by the variance of the outcome variables, as is common the the child development literature. Moreover, using the multiple measures to predict, fewer, underlying dimensions reduces the loss of power resulting from correcting inference for testing multiple conceptually similar outcomes.

Let $\theta^{\mathbf{q}}$ be a vector representing construct q, where q can be child development, preschool quality or the quality of the home environment; $\theta^{\mathbf{q}}$ cannot be directly observed. Then let $m^{j,q}$ be a measure j that is informative, but not perfectly, about $\theta^{\mathbf{q}}$ and let $\epsilon^{j,q}$ be the measurement error associated with how $\theta^{\mathbf{q}}$ maps into $m^{j,q}$. Our approach

corresponds to the treatment having zero effect on *each* child rather than on average, Young (2018) shows that in the case where the sharp null is false randomization-t still performs well in practice and hence we use randomization-t over straightforward randomization inference.

specifies a function that represents the measurement system:

$$m^{j,q} = g^j(\theta^{\mathbf{q}}, \epsilon^{jq}) \tag{3.1}$$

The way we specify the function $g^j()$ depends on the nature of the measurement we have. Given this structure, we can use estimates of the parameters of function in equation (3.1) and a set of measures m^{jq} to obtain estimates of the factors $\theta^{\mathbf{q}}$.

When we have sufficient measures, we assess the dimensionality of the conceptual constructs we consider and estimate underlying latent factors for each dimension. In this section, we discuss our choice of measures, how we score them, how we assess the dimensionality of the underlying constructs and how we then predict esach dimension.

3.1 Child Development

Child development is a multidimensional construct (Cunha et al. 2010; Attanasio et al. 2015; Attanasio et al. 2017) and preschool has been shown to impact various dimensions (Datta Gupta and Simonsen 2010; Berlinski et al. 2009; Chetty et al. 2011; Heckman et al. 2013; Araujo et al. 2016; Kline and Walters 2016). Our child development assessments sought to capture cognition, receptive language, expressive language, school readiness, executive function and socio-emotional development using fifteen different instruments (Table A.2).[17] Preserving the construct validity – the extent to which an instrument measures what it aims to measure – and the reliability or consistency of an instrument when translating and adapting it across languages and cultures is challenging (Peña 2007). Therefore, we selected assessments that had previously been validated for use in Latin American populations; most had previously been used in Colombia (Andrew et al. 2018; Bernal and Fernández 2013).

Trained psychologists, who were blinded to treatment status, carried out eight direct assessments of children's cognition, receptive language, expressive language, school readiness and executive function at endline in the HIs. Five of these instruments, covering concept formation, fluid reasoning, memory for words, rhymes and expressive language, came from the Woodcock-Muñoz-III (WM) tests of cognition and achievement (Schrank et al. 2005) which are Spanish versions of the well-known Woodcock-Johnson tests (Woodcock 1977). To measure receptive language we used the Spanish version of the gold-standard Peabody Picture Vocabulary Test (PPVT) - Test Visual de Imágenes Peabody (TVIP) (Dunn et al. 1986) - and we measured school readiness using the Daberon-II (Danzer

[17] In this section we outline our outcome measures of child development – i.e. those collected at endline. For information about child development measures collected and baseline and used as controls see Section 2.4.

et al. 1991). Inhibitory control, a dimension of executive functioning, was measured using the non-verbal Pencil Tapping Task (Diamond and Taylor 1996). Following test guidelines, two instruments, the Pencil Tapping Task and the Sound Awareness assessment from the WM Achievement battery, were only performed with children above 48 months of age.

Directly assessing socio-emotional development in young children is challenging. We relied on maternal reports, collected as part of the household survey, using the Socio-Emotional Questionnaire in the Ages and Stages Questionnaires (ASQ-SE) (Squires, Bricker, and Twombly 2002). We used all seven subscales of the ASQ-SE, each measuring a different aspect of socio-emotional development: Self-Regulation, Compliance, Communication, Adaptive Functioning, Autonomy, Affect and Interaction with People.[18]

For child development outcomes, we specify the measurement system in equation (3.1) in two steps. We first use an Item Response Theory (IRT) model, which we describe below, to convert a number of binary item responses into a continuous measure, effectively creating a way to score the relevant test. We then use these continuous measures to identify the relevant dimension of the process, estimate a measurement system and obtain estimates of the relevant latent factors.

For thirteen out of fifteen assessments we have item-level data available, i.e. how each child, or mother for the socio-emotional assessments, answered each item in the assessment. For these cases we use a two-parameter IRT model to score the assessments.[19] IRT models whether or not a child answers each item in a given assessment correctly as depending on one underlying latent variable – representing the child's ability in that assessment – and an additive idiosyncratic error which is assumed to be type-I extreme value across items and individuals. The latent ability variable is assumed to be normally distributed with zero mean and unit variance in the population. Items may vary in both their difficulty – how likely is a child of average skill to answer correctly – and their discriminatory power – how sensitive the probability of answering correctly is to the underlying latent ability. IRT scoring proceeds in two stages. In the first, we estimate the difficulty and discrimination of each item from the item responses and assumptions on the distribution of unobservables. In the second we construct assessment scores for each child as the mean of the posterior distribution of the latent ability variable conditional on the child's item responses. See Appendix D.1 for details of the identification and estimation of IRT models.

There are several advantages to using a structural IRT measurement model to score assessments. First, if the

[18] Possible responses to items in the ASQ:SE are 'most of the time', 'sometimes' and 'rarely or never'. Given most respondents chose the option indicating their child had higher levels of socio-emotional development, we code these responses as 1 and the other two as 0.

[19] Two child development assessments, the Woodcock-Muñoz fluid reasoning and the Pencil Tapping Task, simply measure the number of correct responses a child gave. For these two measures we take the count variable as our pre-standardized score.

assumptions of the IRT model are correct then this method uses the information contained within item responses efficiently and allows us to construct estimates of children's skills with the lowest expected error. This is especially relevant when items vary substantially in their difficulty and discrimination power, as may be likely in our situation given the instruments we used were not originally developed for, nor validated in, our population. Second, IRT is well placed to deal with the 'stopping rules' that are features of many of our assessments. These require, for example, that a test is stopped after item x if a child has obtained less than a certain score until that point. With such assessments it is not appropriate to simply use 'naive' raw scores calculated by adding up the number of correct answers since different children attempt different number of questions. Moreover, whether a child attempts a question is often not a smooth function of their performance in the previous questions so there are sharp discontinuities in the distribution of the naive scores at the specified cutoffs. IRT corrects for these problems since in this framework, conditional on her latent ability, whether or not a child attempts a given item is only a function of idiosyncratic errors associated with previous items. The resulting predicted scores are, therefore, no longer discontinuous around the cut-offs.[20]

Our IRT scoring model gives several useful indicators of how well different instruments performed. Well-performing instruments, i.e. those where we can extract a lot of information about the underlying latent skill from the item responses, are those with items with high discriminatory power in the ability range that most children are situated in and where items differ in their difficulty levels. Item Characteristic Curves (ICCs) in Figure B.2, which plot the probability of a child answering each item correctly as a function of their underlying latent ability, show that our measures of receptive language, expressive language, memory for words and school readiness performed well: the ICCs have steep positive slopes (i.e. high discriminatory power) in the range where most children are situated (-3 to 3) and items differ in their difficulties. Our measures of concept formation and sound awareness on the other hand performed poorly, primarily because these assessments were too hard for many children so that many did not progress past the initial few items, leaving very little information. Assessments of socio-emotional skills (subscales of the ASQ:SE) appear substantially less discriminatory with items having flatter ICCs (Figure B.3). The Affect subscale performed especially poorly. We can also estimate the mean reliability for each measure, namely the proportion of the variance of the predicted score that captures true variance in the

[20] Often, an official scoring algorithms is provided; it converts patterns of responses into standardized scores using parameters estimated using a one-parameter IRT model on a norming sample. The validity of these, however, depends crucially on how similar the performance of items in the norming sample is to the study sample. This is an issue for work with specific sub-groups or contexts not covered by the norming sample. For example, the norming sample for the Woodcock-Muñoz test consists of 1,413 Spanish speaking children from the USA, 6 Latin American countries and Spain (Schrank et al. 2005) who are likely to differ substantially from our sample of children in Colombia.

underlying construct. Again, the estimated mean reliabilities suggest that concept formation and sound awareness performed substantially worse than other directly assessed measures and that assessments of socio-emotional skills were less informative than measures of cognition, language and school readiness (Tables A.3 and A.5).

We standardize the IRT scores for age non-parametrically. We use locally weighted regressions to estimate the mean and variance in the control group at each age and then use these to create z-scores, for each instrument (see Appendix D.2 for details). We label these age-standardized scores m_c^j, for child c's score for measure j. The scaling of these scores, which have zero mean and unit variance in the control group at each age, implies that the magnitude of treatment effects can be interpreted relative to the age-specific standard deviation of the control group. We exclude from the analysis observations with standardized values lower than 3 standard deviations below the mean (<-3SD), since we consider this to be an indication of either the assessment being carried out incorrectly or of potential disability.

Given the challenges of adapting child development assessments across contexts (Peña 2007) we check our measures pass basic tests of internal validity. Six of our eight direct measures of child skills (measuring cognition, language, school readiness and executive function) are strongly correlated with age, baseline child development and household wealth in the expected direction (Table A.3) and are strongly positively correlated with one another (Table A.4). The two measures that presented problems in the IRT procedure (concept formation and sound awareness) are the ones that are not strongly correlated with these variables. Maternal report measures of socio-emotional development show lower correlations with age, baseline socio-emotional development, household wealth and maternal education (Table A.5) and with one another (Table A.6) than the direct-assessment measures. This could be a feature of socio-emotional skills or a sign that the maternal-report measures are of poorer quality.

Based on a well-established literature on the different domains of early child development we pre-specified three, not mutually exclusive, groups of measures: (i) cognitive development, language development and school readiness (which we refer to as CLS), (ii) pre-literacy skills and (iii) socio-emotional skills.[21] Within each group, we assess the dimensionality of the contained measures - the extent to which each represents a separate construct or whether they can be summarized by fewer constructs. For each of these three groups of measures, of size J, we estimate a linear exploratory factor analysis model:

$$\boldsymbol{m}_c = \boldsymbol{\theta}_c \Gamma + \xi_c$$

[21] Pre-specified in pre-analysis plan held at AEA trial registry (AEARCTR-0001246). CLS includes Receptive Language, Concept formation, Fluid reasoning , Expressive language, Memory for words, School readiness and Inhibitory control. Pre-literacy skills include Receptive language, Expressive Language, Memory for words, and Rhymes. Socio-emotional skills includes the seven subscales of the ASQ-SE.

where $\boldsymbol{m}_c = [m_c^1, m_c^2 ... m_c^J]$ is the vector of J age-standardized scores, where $\boldsymbol{\theta}_c = [\theta_c^1, ... \theta_c^k]$ is a vector of underlying latent factors of dimension $k < J$, Γ is the $k \times J$ matrix of factor loadings and ξ_c is a vector of errors which are assumed to be uncorrelated with $\boldsymbol{\theta}_c$ and independent from one another. The un-rotated factor loadings are pinned down by assuming the underlying latent factors are uncorrelated and normalizing the scale of the variance to $V(\boldsymbol{\theta}_c) = \mathbf{I}$. We estimate factor loadings using the iterated principal-factor method and follow the standard method of retaining all factors with an eigenvalue greater than one (Kaiser 1960). In this analysis we drop three measures that have factor loadings of less than 0.4, as specified in the pre-analysis plan: concept formation, sound awareness and affect. These measures all performed poorly in IRT analysis and exhibited poor internal validity. We use the estimated factor loadings to predict values of these underlying latent factors for each child using the regression method (Thomson 1939).

All three groupings result in a single factor being retained (Tables A.7 and A.8) leaving us with three dimensions of child development on which to assess the program's impacts: Cognitive, Language and School Readiness (CLS), Pre-literacy and Socio-Emotional skills.

3.2 Preschool Quality

In addition to extensive measurement of child development, a distinguishing feature of this study is in-depth assessment of the quantity and quality of teachers' classroom activities.

We collected detailed teacher-reported data on the type and frequency of activities teachers had performed in the classroom over the week prior to the interview using the Teacher Survey of Early Education Quality (Hallam, Rous, Riley-Ayers, and Epstein 2011). In particular, we asked whether the teacher had performed each of 36 different activities and, if so, on how many days. We split the activities into "Learning and Development Activities", such as reading stories, teaching skills, storytelling and singing, and "Personal Care Activities" related to basic care of children such as changing nappies, brushing teeth and washing hands, naps and feeding routines. We calculate measures for the two activity types as the sum, across all relevant activities, of the number of days on which the teachers performed the activity.

The pedagogical training targeted the quality, in addition to the quantity, of classroom activities. We measured quality by direct observation using the Early Childhood Environmental Rating Scale - Revised (ECERS-R) (Harms, Clifford, and Cryer 1998). The ECERS-R measures the quality of the learning environment across six dimensions

– Space and Furnishings, Personal Care Routines, Language and Literacy, Learning Activities, Interaction and Program Structure. The instrument has been used extensively across a wide range of cultural and economic contexts and has been shown to be predictive of child gains across cognitive (Peisner-Feinberg et al. 2001; Burchinal et al. 2000) and social-emotional skills (Sylva et al. 2006).

The ECERS-R was carried out by psychologists, who were trained for three weeks, and each observation lasted at least half a school day. Due to logistical and budgetary constraints, we only conducted ECERS-R in 172 of the 847 classrooms in our sample.[22]

The ECERS-R is comprised of 43 individual items, each measuring a different aspect of quality, for example 'encouraging children to communicate'. Each item is formed of around 10 sub-items grouped under the headings 'inadequate', 'minimal', 'good' and 'excellent' to which the observer must answer 'true' or 'false'. We followed the official administration procedure, which unfortunately turned out to be poorly suited to our context due to stopping rules which resulted in a high number of non-random missing values for items in the 'minimal', 'good' and 'excellent' categories (See Appendix D.3 for details). We therefore only use items from the 'inadequate' category in our analysis. While this overcomes the challenge posed by missing data, it implies that the sub-items that make up our quality measures are informative on the absence of poor practices rather than the presence of good ones.

We group together all sub-items in the Personal Care Routines, Language and Literacy, Learning Activities, Interaction and Program Structure subscales to construct a measure of *process* quality while using items in the Space and Furnishings subscale to construct a separate scale.[23] We drop items with very low variance (< 0.10) and predict the underlying latent process quality and space and furnishings for every classroom using a two-parameter IRT model, as described above.

ICCs for both measures show that while the majority of the items perform well and are informative about underlying classroom quality, they tend to be more informative at lower levels of underlying quality meaning the measures have more discriminatory power in identifying changes in quality in poor quality classrooms than high

[22] The subsample was chosen as follows. At baseline, we randomly chose 216 classrooms attended by study children in 54 HI's selected randomly, stratifying by city, in which to measure classroom quality using either the ECERS-R (suitable for classrooms with children over two years of age, 60% of classrooms) or ITERS-R (corresponding assessment for classes of children aged 0-2, 40% of classrooms). At follow-up, we had sufficient budget to collect observations on 211 classrooms in 54 centers. We chose half these classrooms to be the same classrooms we had observed at baseline (randomly chosen) and the other half to be classrooms attended by children in the sample at follow-up (since study children had moved on from their baseline classrooms). This resulted in observations in 172 classrooms with children older than two years where we carried out the ECERS-R and 39 classrooms with children aged 0-2 where we carried out the ITERS-R. Given the small size of the ITERS-R sample at endline and given that there are no common items with the ECERS-R that would allow us to create a common measure, we drop these classrooms from our classroom analysis sample.

[23] To increase the sample size for estimating IRT parameters we pool ECERS-R measures from baseline and end-line giving a total sample of 296 observations. We drop items that were found to be preventing model convergence. These were items where the estimated discriminatory parameter after 5 iterations was negative implying that the item was functioning poorly.

quality classrooms (Figure B.4).

3.3 Quality of the Home Environment

We assessed the quality of the home environment in order to investigate whether the programs, either directly or indirectly, impacted parental behavior. We used an adapted version of UNICEF's Family Care Indicators (FCIs) which were themselves adapted from the Home Observation for Measurement of the Environment (Hamadani, Tofail, Hilaly, Huda, Engle, and Grantham-McGregor 2010). We focus on the two dimensions measured by the FCI that have previously been shown to be predictive of child development – the variety of play materials and play activities (Hamadani, Tofail, Hilaly, Huda, Engle, and Grantham-McGregor 2010). Compared to the original FCI, which asks parents yes/no questions about the play activities the child engaged in over the past three days and the play materials the child played with, we extended both subscales due to concerns that binary items are insufficient to pick up adequate variation in the quality of the home environment outside of contexts of extreme poverty (Culhane, Cunha, Elo, and Pham 2016). In particular, for the play materials subscale we collected information on the number of each type of toy that a child played with.[24] For the play activities subscale, we distinguish between play activities carried out with the mother and with the father and separately recorded those carried out over the last five weekdays and over the last weekend.[25] For each we recorded the number of times or the amount of time taken on the activity (rather than just whether it had been done).

With the three domains of, first, play materials, second, play activities with the mother and, third, play activities with the father we estimate an exploratory factor analysis model, as described by equation 3, to assess the dimensionality of our multiple continuous measures. We find that, for each domain, all measures load onto a single underlying latent factor (Tables A.9, A.10 and A.11) which we then predict for each child.

4 Results

In this section, we report our results. We start with the estimates of the impacts of HIM and HIM+FE programs on child development followed by analysis of impact heterogeneity; we then move onto an analysis of possible

[24] As in the original FCI we asked about play materials across the following categories: Things which make/play music; Things for drawing/writing; Picture books for children (not school-books); Things meant for stacking, constructing, building (blocks); Things for moving around (balls, bats, etc.); Toys for learning shapes and colours; Things for pretending (dolls, tea-set, etc.).

[25] As in the original FCI we asked about play activities across the following categories: Read books or look at picture-books with child; Tell stories to child; Sing songs with child; Take child outside home place; Play with the child with toys; Spend time with child in naming things, counting, drawing.

mechanisms that could explain the estimated impacts.

4.1 Effects on Child Development

The results show a stark difference between the impacts of the HIM and HIM+FE programs (Table 2). HIM, the government quality improvement program, did not have the positive impact on child development that it aimed to achieve. There is even some indication that HIM may have had adverse effects relative to the 'business-as-usual' control preschools. The estimated HIM treatment effect is negative for all three dimensions of child development (see Table 2 and Table 3) with quite a large negative treatment effect for pre-literacy skills (-10% of a standard deviation) which is approaching statistical significance ($p = 0.141$).[26][27]

In contrast, the enhanced program (HIM+FE) significantly improved children's cognition, language and school readiness by 15% of a standard deviation relative to the 'business-as-usual' control preschools. The final row of Table 2 shows that the addition of the FE component to the HIM upgrade, the current status-quo, increased the effectiveness of the government quality enhancement program by raising children's cognitive, language and school readiness score by 18% of a standard deviation and pre-literacy score by nearly 15% of a standard deviation. These effects remain highly statistically significant once p-values are corrected for testing hypotheses on the two outcomes simultaneously. Examining effects of HIM+FE on the individual measures of child development, presented in Table A.12, suggests that this effect is driven by improvements in performance on tests on fluid reasoning and expressive language. The stark difference in the effectiveness of the government quality improvement program and the quality improvement program combined with the FE program suggests that the FE component is crucial.

The effects of the HIM+FE program, relative to the 'business as usual' control, come through most strongly for the older children in the sample and are not significantly different from zero once children under the age of 48 months at follow-up are included in the sample although the difference between HIM+FE and HIM continues to be significant (Table 2, final column).[28] This could either be because the intervention was genuinely more effective

[26] While all tables report p-values calculated through randomization inference (randomization-t) in addition to through a block bootstrap, we notice that these are typically very similar. Thus we report statistical significance on the basis of p-values calculated through the block bootstrap.

[27] While baseline measures were taken after HIM had formally begun, we see no very short term impacts (see baseline balance table). If any very short term impacts, in the same direction as impacts at endline, were present then the inclusion of baseline controls would attenuate estimated impacts. Our results are robust to excluding child development controls from baseline. Results on the impact of FE over and above HIM would be unaffected since FE began after baseline.

[28] As described, the Pencil Tapping Test (PTT) is not suitable for children under the age of 48 months. Therefore, the complete cognitive, language and school readiness measure can only be constructed for children older than 48 months at follow-up (1,071 out of 1,819). Estimates in the first and last columns of Table 2 show that while HIM+FE had an impact on the older sub-sample for whom the complete measures can be constructed, there are no impacts once the younger children are included on the measure which excludes the PTT. Column 3 further shows that this difference is not being driven by the exclusion of the PTT assessment since the significance and size of HIM+FE impact on the older sub-sample does not change once PTT is excluded.

Table 2: Impact of HIM and FE on Cognition, Language and School Readiness and Pre-Literacy

	Pre-Specified Analysis		***Exploratory Analysis***	
	Cognitive, Language and School Readiness (CLS)	Pre-literacy Skills	CLS (exc. PTT, limited sample)	CLS (exc. PTT, full sample)
	(1)	(2)	(3)	(4)
HIM only	-0.030	-0.100	-0.041	-0.066
	(0.079)	(0.068)	(0.079)	(0.066)
	$[p = 0.693, p^{FW} = 0.693]$	$[p = 0.141, p^{FW} = 0.200]$	$[p = 0.606]$	$[p = 0.317]$
	$\{p = 0.682\}$	$\{p = 0.124\}$	$\{p = 0.585\}$	$\{p = 0.297\}$
HIM+FE	0.151**	0.045	0.148**	0.066
	(0.074)	(0.064)	(0.075)	(0.064)
	$[p = 0.039, p^{FW} = 0.066]$	$[p = 0.481, p^{FW} = 0.481]$	$[p = 0.048]$	$[p = 0.310]$
	$\{p = 0.045\}$	$\{p = 0.465\}$	$\{p = 0.055\}$	$\{p = 0.277\}$
HIM + FE vs. HIM	0.182***	0.145***	0.189***	0.132**
	(0.066)	(0.052)	(0.067)	(0.055)
	$[p = 0.005, p^{FW} = 0.010]$	$[p = 0.006, p^{FW} = 0.006]$	$[p = 0.005]$	$[p = 0.016]$
	$\{p = 0.017\}$	$\{p = 0.016\}$	$\{p = 0.014\}$	$\{p = 0.035\}$
N	1071	1819	1071	1819

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$ using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across our two pre-specified child development outcomes, in columns (1) and (2). Columns (3) and (4) present additional exploratory analysis and thus we do not correct for multiple testing here. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, tester effects and baseline scores for MacArthur Bates CDI 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 4. Age effects removed from standardized scores prior to factor construction. Each factor is constructed using the following standardized scores: (i) Cog, Lang and Sch (all measures): WM12, WM14, WM17, TVIP, DAB and PTT, (ii) Cog, Lang and Sch (exc. PTT): WM12, WM17, TVIP, WM14 and DAB, and (iii) Pre-literacy skills: WM14, WM17 and TVIP. See Table A.2 for details of measures and factors.

Table 3: Impact of HIM and FE on socio-emotional development

	Socio-Emotional Skills
	(1)
HIM only	-0.014
	(0.087)
	$[p = 0.874]$
	$\{p = 0.851\}$
HIM+FE	0.034
	(0.089)
	$[p = 0.710]$
	$\{p = 0.690\}$
HIM + FE vs. HIM	0.047
	(0.082)
	$[p = 0.562]$
	$\{p = 0.542\}$
N	1826

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [$bootstrap$] and (ii) $\{randomization - t\}$. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, tester effects and baseline scores for ASQ:SE. Factor scaled to have a mean of zero and standard deviation of one in the control group. Factor constructed as described in section 4. Age effects removed from standardized scores prior to factor construction. Factor is constructed using the following standardized scores: Self Regulation, Compliance, Communication, Adaptive Functioning, Autonomy, Affect, Interaction with People.

for older children or because our tests had higher discriminatory power for this sub-sample.

As with cognitive development, Table 3 shows that the additional resources provided through the government quality improvement program had no impact on child socio-emotional development. In contrast to findings on cognitive outcomes, however, the addition of the FE component did little to change this (final row, Table 3).

4.2 Heterogeneity by Household Wealth

Several studies from high income countries show that children from disadvantaged households benefit more from access to child care than children from better off backgrounds (Cornelissen, Dustmann, Raute, and Schönberg 2018; Felfe and Lalive 2018; Havnes and Mogstad 2015). We thus assess heterogeneity by household wealth. In interpreting these results, however, it should be noted that our sample consists of children of low SES *working* parents, who, because they are working, are not as poor as those typically eligible for public programs in Colombia.[29] For example, the households in our sample had an average monthly income of roughly 1.3 million COP (US$285), equivalent to approximately two minimum wages in 2013, and average parental education was relatively high at 12 years.

[29] The government targets these children because with two working parents and a relatively low income household, they are at especially high risk of receiving inadequate childcare.

Table 4: Heterogeneity by Child and Household Characteristics

	CLS	Pre-literacy Skills	CLS	Pre-literacy Skills
	(1)	(2)	(3)	(4)
HIM only	0.040 (0.108) $[p = 0.706, p^{FW} = 0.706]$ $\{p = 0.690\}$	-0.068 (0.083) $[p = 0.416, p^{FW} = 0.712]$ $\{p = 0.409\}$	-0.002 (0.123) $[p = 0.986, p^{FW} = 0.986]$ $\{p = 0.988\}$	-0.082 (0.098) $[p = 0.403, p^{FW} = 0.725]$ $\{p = 0.381\}$
HIM only X Higher BL Dev			-0.070 (0.140) $[p = 0.613, p^{FW} = 0.895]$ $\{p = 0.600\}$	-0.036 (0.111) $[p = 0.744, p^{FW} = 0.925]$ $\{p = 0.740\}$
HIM only X Wealthier	-0.128 (0.126) $[p = 0.309, p^{FW} = 0.621]$ $\{p = 0.230\}$	-0.052 (0.101) $[p = 0.606, p^{FW} = 0.832]$ $\{p = 0.587\}$		
HIM+FE	0.302*** (0.094) $[p = 0.001, p^{FW} = 0.004]$ $\{p = 0.004\}$	0.158* (0.088) $[p = 0.069, p^{FW} = 0.069]$ $\{p = 0.072\}$	0.237** (0.098) $[p = 0.016, p^{FW} = 0.048]$ $\{p = 0.038\}$	0.096 (0.084) $[p = 0.257, p^{FW} = 0.393]$ $\{p = 0.264\}$
HIM+FE X Higher BL Dev			-0.172 (0.130) $[p = 0.178, p^{FW} = 0.358]$ $\{p = 0.222\}$	-0.103 (0.108) $[p = 0.336, p^{FW} = 0.336]$ $\{p = 0.348\}$
HIM+FE X Wealthier	-0.286*** (0.111) $[p = 0.009, p^{FW} = 0.022]$ $\{p = 0.022\}$	-0.219** (0.100) $[p = 0.031, p^{FW} = 0.052]$ $\{p = 0.044\}$		
HIM+FE vs. HIM only	0.262*** (0.088) p=0.003 $p^{FW} = 0.007$ $\{p = 0.009\}$	0.226*** (0.064) p=0.001 $p^{FW} = 0.002$ $\{p = 0.005\}$	0.238** (0.113) p=0.033 $p^{FW} = 0.070$ $\{p = 0.035\}$	0.178** (0.082) p=0.029 $p^{FW} = 0.073$ $\{p = 0.062\}$
HIM+FE vs. HIM only X Higher BL Dev			-0.102 (0.152) p=0.500 $p^{FW} = 0.649$ $\{p = 0.500\}$	-0.068 (0.116) p=0.552 $p^{FW} = 0.552$ $\{p = 0.571\}$
HIM+FE vs. HIM only X Wealthier	-0.158 (0.124) p=0.204 $p^{FW} = 0.204$ $\{p = 0.192\}$	-0.167* (0.087) p=0.053 $p^{FW} = 0.087$ $\{p = 0.064\}$		
N	1071	1819	1071	1819

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across our two pre-specified child development outcomes and, for each of the two heterogeneity analyses separately, across the two subgroups. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, tester effects and baseline scores for MacArthur Bates CDI 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 4. Age effects removed from standardized scores prior to factor construction. Each factor is constructed using the following standardized scores: (i) Cog, Lang and Sch (all measures): WM12, WM14, WM17, TVIP, DAB and PTT, (ii) Cog, Lang and Sch (exc. PTT): WM12, WM17, TVIP, WM14 and DAB, and (iii) Pre-literacy skills: WM14, WM17 and TVIP. See Table A.2 for details of measures and factors. Higher BL development implies child had above median value of ASQ-III factor score at baseline. Wealthier implies child's household had above median value of household asset index at baseline.

Heterogeneity analysis by household wealth suggests that in this setting too the average treatment effect of HIM+FE is driven by effects on the poorer children. Breaking the sample down into two groups – children from households with below and above median wealth – shows that HIM+FE, relative to the control, raised Cognitive, Language and School Readiness skills by 30% of a standard deviation and pre-literacy skills by 16% of a standard deviation among the poorer children; it had no significant impacts on the better-off children (Table 4).

Generally, the finding that children from poorer households benefit more from center-based care than children from better off households might be explained by (i) the fact that the counterfactual way in which they would spend their time (at home) is less productive than that for better off children; and (ii) if, holding counterfactual activities constant, children with lower developmental levels have more to gain from being in a stimulating environment. Since we are studying changes in quality rather than intensity of exposure to preschool, the latter explanation appears more likely here. And, indeed, results in Table 4 show that HIM+FE impacts were only significant for children with lower baseline levels of development.[30] It continues to be the case that the government quality improvement program, HIM, has no impacts even on the more disadvantaged children.

4.3 Potential Mechanisms

To understand the mechanisms that generate both the null, bordering on negative, impacts of HIM relative to the control and the positive impacts of HIM+FE, we examine how both interventions affect the quantity and quality of classroom activities and the quality of children's home environment.

Reduced-form correlations suggest that the way teachers spend their time matters. In Table 5, we report the correlation between teachers' self-reported classroom activities, averaged at the center level, and child development.[31] There is a significant positive association between the number of learning and development activities teachers in the preschool carry out and our main measure of child development at endline. However, no correlation can be seen between the number of personal care based activities and child development. We note that the estimates, perhaps not surprisingly, are not very precise: when we limit the analysis to the control sample (which is one third of the size), the estimated coefficients do not change substantially, but become statistically insignificant.

We find that in response to HIM, the government improvement program, teachers significantly reduced the

[30] As measured by five subscales of an extended version of the ASQ-3 - Communication, Gross Motor, Problem Solving, Personal-Social and Fine Motor (Squires, Bricker, and Twombly 2009); MacArthur Bates Communicative Development Inventories (Jackson-Maldonado et al. 2013; Jackson-Maldonado et al. 2003); to measure language development; and ASQ:SE (Squires, Bricker, and Twombly 2002) to measure socio-emotional development.

[31] We average teachers' activities at the preschool level as most children have been taught by more than one teacher during their time at the HI.

Table 5: Correlations between Child Development and Teacher Reported Classroom Activities

	(1)	(2)	(3)	(4)	(5)	(6)
Learning and Development Activities	0.0912**		0.123*	0.160**	0.179	0.175
	(0.0429)		(0.0693)	(0.0674)	(0.176)	(0.154)
Personal Care Activities		0.0235	-0.0444	-0.0504	0.0279	0.00148
		(0.0415)	(0.0587)	(0.0535)	(0.104)	(0.098)
Observations	1071	1071	1071	1071	350	350
BL Child Development	No	No	No	Yes	No	Yes
Sample	All	All	All	All	Control	Control

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (clustered at HI level) in parentheses. Table presents OLS regression coefficients for regression of Cognitive, Language and School Readiness (CLS) child development factor on teachers' involvement in (1) Learning and Development Activities and (2) Personal Care Activities. Construction of all measures are described in Section 3. Measures of teachers' involvement in classroom activities averaged across all teachers in the HI. All regressions control for city effects and child gender. Regressions (4) and (6) additionally control for baseline child development as measured by the MacArthur Bates CDI and each sub-scale of the ASQ-III.

number of, and thus presumably the time spent on, both learning and development activities and personal care activities (Table 6). Of the new staff hired, only the classroom assistants were added to staffing of the classes[32]. The reduction in personal care activities is expected and a direct result of the design of the program: the classroom assistants were supposed to be primarily in charge of receiving kids in the morning, transitions, personal care routines, preparation of materials, and helping the teacher during the activities. What is surprising is that the number of learning activities decreased. The available data do not allow us to check whether the new staff are substituting for teachers also in these activities but we can conclude that teachers are doing less of them. The associations between teachers' time allocation and child development (Table 5) suggest that while the reduction in time spent on personal care based activities is unlikely to have detrimental impacts on child development, the reduction in time spent on learning and development activities may.

These findings are somewhat similar to findings from the Tennessee's Project STAR which looked at the impacts of reducing class sizes by one third and adding a teaching aide to regular size classrooms on children's performance. That study found that while reducing class sizes improved children's performance, especially in kindergarten classes (Hanushek 1999), the addition of teaching aides had no impact (Mosteller 1995; Gerber, Finn, Achilles, and Boyd-Zaharias 2001). These findings are consistent with our results and suggest that while the intensity and quality of time from qualified teachers is an important determinant of children's performance, the introduction of less qualified staff into the classroom may have very different impacts.

The unintended reduction in what are likely to be productive classroom activities in response to the government

[32] The socio-emotional and nutrition experts were hired to spend one-on-one time with the children as well as advise teachers and parents.

Table 6: Impacts on Teacher Reported Classroom Activities

	Learning and Development Activities	Personal Care Activities	Reading Activities
	(1)	(2)	(3)
HIM only	-0.343**	-0.340**	-0.094
	(0.138)	(0.151)	(0.116)
	$[p = 0.012, p^{FW} = 0.023]$	$[p = 0.027, p^{FW} = 0.027]$	$[p = 0.411]$
	$\{p = 0.008\}$	$\{p = 0.011\}$	$\{p = 0.332\}$
HIM+FE	-0.113	-0.331**	-0.052
	(0.113)	(0.157)	(0.088)
	$[p = 0.300, p^{FW} = 0.300]$	$[p = 0.016, p^{FW} = 0.030]$	$[p = 0.654]$
	$\{p = 0.316\}$	$\{p = 0.010\}$	$\{p = 0.586\}$
HIM + FE vs. HIM	0.231**	0.009	0.042
	(0.110)	(0.131)	(0.115)
	$[p = 0.040, p^{FW} = 0.075]$	$[p = 0.951, p^{FW} = 0.951]$	$[p = 0.638]$
	$\{p = 0.008\}$	$\{p = 0.950\}$	$\{p = 0.657\}$
N	847	847	847

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across our two categories of classroom activities, in columns (1) and (2). Column (3) presents additional exploratory analysis and thus we do not correct for multiple testing here. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for city effects, tester effects and average baseline scores for MacArthur Bates CDI at the HI level. All measures scaled to have a mean of zero and standard deviation of one in the control group.

quality improvement program seems to be offset by the addition of the FE enhancement which emphasized the importance of prioritizing learning and development activities over personal care routines and provided coaching on productive strategies for implementing these activities. The results in Table 6 suggest that teachers in the HIM+FE arm reduced the time they spent on the personal care activities, those activities that the classroom assistants are well qualified to take over, relative to the pure controls, but not the learning and development activities. Relative to the current status-quo (HIM), the addition of the FE enhancements substantially increases the time teachers spend on learning and productive activities (Table 6).

In addition to teacher reported measures, we also have direct observations of the quality of classroom activities from the ECERS-R instrument, which captures the quality of the infrastructure as well as teaching and care provision within preschools. These measures are expensive to collect, so they are only available for subset of 172 of the 847 classrooms in the 120 preschools in our sample. However, these data can be used to assess impacts of the interventions on teachers' practices and classroom processes. Estimates of treatment effects on these measures, presented in Table 7, suggest that the FE enhancement did not only offset the negative impacts of HIM on the amount of time teachers spend on "Learning and Development Activities" with the kids, but also had a positive

Table 7: Impact on Directly Observed Classroom Quality (ECERS-R)

	Process Quality	Space and Furnishings
	(1)	(2)
HIM only	-0.022	0.117
	(0.244)	(0.226)
	$[p = 0.933, p^{FW} = 0.933]$	$[p = 0.612, p^{FW} = 0.848]$
	$\{p = 0.933\}$	$\{p = 0.564\}$
HIM+FE	0.412*	0.249
	(0.225)	(0.215)
	$[p = 0.065, p^{FW} = 0.122]$	$[p = 0.242, p^{FW} = 0.242]$
	$\{p = 0.058\}$	$\{p = 0.207\}$
HIM + FE vs. HIM	0.434*	0.132
	(0.241)	(0.198)
	$[p = 0.072, p^{FW} = 0.140]$	$[p = 0.503, p^{FW} = 0.503]$
	$\{p = 0.131\}$	$\{p = 0.491\}$
N	172	172

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across our two categories of preschool quality, in columns (1) and (2). Bootstrapped statistics use block-bootstraps, resampling HIs with replacement (since not all HIs are covered we do not resample triplets), and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for city effects and average baseline scores for MacArthur Bates CDI at the center level. Both measures scaled to have a mean of zero and standard deviation of one in the control group. Both measures constructed as described in section 4.

effect on the quality of this time, as evidenced by a significant increase of around two fifths of a standard deviation in "process" quality in the HIM+FE arm relative to both the pure control and the HIM arms[33]. We see no change in the quality of "space and furnishings" in either arm.[34]

As well as the pedagogical training component, the FE enhancement contained a reading program which targeted parents; a change in quality of the home environment is, therefore, another potential mechanism for the positive impact that the addition of the FE component to HIM had on child development. We find no evidence of this, however; results in Table 8 show that neither HIM nor HIM+FE had any impact on any of the domains of quality of the home learning environment measured using the FCI. Even when looking at specific activities within the sub-scales, we see no impacts on the time that parents spent reading with the child – the activity directly targeted by the FE reading program. We also see no change in the number of times that teachers read stories to their kids

[33] It should be noted that we only have center quality measure for a sub-sample of classes which reduces the precision of our impact estimates and limits options for more in-depth investigation of this potential mechanism.

[34] We do not report results using individual children outcomes as the limited number of observations yields very imprecise estimates. Results are available on request.

in a week (Table 8). These results suggest that the reading program is unlikely to be driving the effect of the FE enhancement on child development.

The lack of impacts on the home environment and reading practices in the class (targeted by the reading program component of FE) combined with an improvement in classroom process quality (targeted by the pedagogical training program component of FE) suggest that the addition of pedagogical training is what is driving the significant impact that the FE enhancement has on increasing the efficacy of the government quality improvement program through (1) offsetting the negative effect that the government program has on the time teachers spend on play and learning activities with the kids in the classroom; and (2) improving the quality of what teachers do with that time. This hypothesis is further supported by evidence that the FE enhancement appears to have been more effective in the smaller centers (Table A.14), where, on average, a higher proportion of teachers (one half vs. one quarter; Table A.13) received the training.[35][36].

[35] This pattern by center size was driven by the fact that only up to three teachers per HI were supposed to participate in the training.

[36] FE were only able to provide us with very partial data on teacher attendance which do not allow us to link all teachers in our data in the HIM+FE arm to data on attendance of the pedagogical training program. However, partially linked data combined with self-reported data from teachers on whether they attended training suggest that a significantly higher proportion of teachers in the smaller centers were trained.

Table 8: Impact of HIM and FE on Learning Environment in the Home

	Summary Indices			*Reading Activities and Materials*		
	Play Materials	Play Activities with Mother	Play Activities with Father	Number of Storybooks	Mins Reading with Mother	Mins Reading with Father
	(1)	(2)	(3)	(4)	(5)	(6)
HIM only	0.020 (0.078) $[p = 0.806, p^{FW} = 0.806]$ $\{p = 0.763\}$	-0.067 (0.078) $[p = 0.384, p^{FW} = 0.617]$ $\{p = 0.368\}$	-0.098 (0.074) $[p = 0.182, p^{FW} = 0.416]$ $\{p = 0.148\}$	-0.050 (0.247) $[p = 0.845, p^{FW} = 0.845]$ $\{p = 0.821\}$	-4.554 (10.453) $[p = 0.657, p^{FW} = 0.886]$ $\{p = 0.618\}$	-6.495 (6.793) $[p = 0.342, p^{FW} = 0.691]$ $\{p = 0.318\}$
HIM+FE	0.109 (0.074) $[p = 0.142, p^{FW} = 0.354]$ $\{p = 0.120\}$	0.064 (0.065) $[p = 0.316, p^{FW} = 0.512]$ $\{p = 0.356\}$	-0.041 (0.062) $[p = 0.500, p^{FW} = 0.500]$ $\{p = 0.537\}$	0.080 (0.219) $[p = 0.720, p^{FW} = 0.918]$ $\{p = 0.694\}$	5.380 (9.064) $[p = 0.556, p^{FW} = 0.904]$ $\{p = 0.555\}$	-1.398 (7.145) $[p = 0.835, p^{FW} = 0.835]$ $\{p = 0.847\}$
HIM + FE vs. HIM	0.089 (0.070) $[p = 0.116, p^{FW} = 0.289]$ $\{p = 0.208\}$	0.131 (0.084) $[p = 0.116, p^{FW} = 0.289]$ $\{p = 0.088\}$	0.057 (0.084) $[p = 0.489, p^{FW} = 0.489]$ $\{p = 0.442\}$	0.130 (0.205) $[p = 0.521, p^{FW} = 0.521]$ $\{p = 0.470\}$	9.933 (8.388) $[p = 0.230, p^{FW} = 0.528]$ $\{p = 0.220\}$	5.098 (6.522) $[p = 0.425, p^{FW} = 0.662]$ $\{p = 0.393\}$
N	1848	1827	1447	1848	1827	1447

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing separately across summary outcomes related to the home environment, in columns (1)-(3), and across outcomes specifically related to reading, in columns (4)-(6). Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, interviewer effects and baseline scores for the play materials and play activities with mother subscales. All factors scaled to have a mean of zero and standard deviation of one in the control group.

5 Conclusions

In this paper we show, within one institutional setting, that different approaches to improving the quality of early years education can have very different impacts. Our results suggest that providing preschools with additional resources for hiring and teaching materials without any training on best practices in the classroom is an ineffective and potentially even detrimental strategy. Complementing such resources with pedagogical training and coaching, on the other hand, can have significant positive impacts on child development, especially for more disadvantaged children. Self-reported data suggest that without training on best practices, teachers respond to having more staff and materials by reducing the time that they spend with the children, including on activities that are important for child development. In contrast, the addition of pedagogical training prevented any reduction in time teachers' spent on learning and development focused activities and increased the observed quality of instruction in the class.

At around 15% of a standard deviation, the estimated positive impact of the enhanced government quality improvement program on the main cognitive development outcome measure corresponds to 15% of the achievement gap between children in the top and bottom wealth quintiles in Colombia by age 5 (Bernal, Martínez, and Quintero 2015). Heterogeneity analysis shows that the impact is higher on the more disadvantaged children, for whom the effect size is much larger at 30% of a standard deviation. To the extent that credible comparisons can be made between intervention evaluations with different measures of child development and in different populations, the effects we estimate of the FE quality improvement program are in the ballpark of other studies which look at effects of children accessing center-based care in Colombia (Nores, Bernal, and Barnett 2019) and other Latin American countries (Noboa-Hidalgo and Urzúa 2012; Berlinski et al. 2008; Bernal and Ramírez 2019; Bernal and Fernández 2013; Behrman et al. 2014). There is little to guide extrapolation of how these short-run impacts might map onto long-run outcomes of children in Colombia. However, evidence from further afield, such as evaluations of Head Start in the US, suggests that programs which achieved short-run effects of similar magnitude can have wide ranging and persistent positive long-run effects (Garces, Thomas, and Currie 2002; Deming 2009).

The immediate policy implication of our findings is that the current strategy of the government of Colombia for improving quality of care in its *Hogares Infantiles* preschools through scaling up the HIM program may be doing more harm than good. However, the results also offer guidance on an effective improvement strategy: the addition of a pedagogical training component following the blue-print of the program designed by the Colombian National University for *Fundación Exito* in this study. Ongoing costs of the current government upgrade are $300 per child

per year, raising the total cost of provision to $1300 per child per year. Our results suggest that the impacts of the government program on child development could be increased from none (possibly negative) to as much as 30% of a standard deviation for the more disadvantaged children attending these centers by complimenting the hiring of new staff with appropriate teacher training and coaching.

A back-of-the-envelope calculation of the costs of scaling-up the pedagogical training component of the FE program, the component that we argue is key to its impacts on children's development suggest that scaling to new centers would require an upfront investment of $5827 per HI. However, once this initial investment had been made, maintaining the proportion of trained teachers through training new teachers and providing less-intensive refresher training to previously-trained teachers would cost $2206 per HI per year, or $13 per child per year. This additional cost thus represents 1% of the current amount of $1300 allocated to HIs yearly for each child. See Appendix E for details of calculations.

The study design does not allow us to say anything about the stand-alone benefits of the pedagogical program in the absence of the government's HIM upgrade. While it is probable that a minimum level of physical resources and staffing may be needed in order for programs focusing on teaching practices to be effective, we cannot tell whether the HIs in this study had already surpassed such a level before the HIM upgrades. Thus we cannot tell whether FE would have been effective even if the government had not implemented HIM or whether it is the interaction between more resources and know-how which is driving the positive impacts in the HIM+FE arm. Nevertheless, we note that recent overviews suggest that programs that directly target children's day to day classroom experience have been successful even in contexts with minimal prior resources (Kremer, Brannen, and Glennerster 2013; Murnane and Ganimian 2014).

More broadly, this study suggests that widespread reliance on easily observable indicators such as teacher-child ratios and availability of play materials to judge overall quality of early years education and target investment may be misguided. Stand-alone improvements in these factors are unlikely to result in the sorts of short and long-run improvements to human capital that some investments in high quality early childhood interventions have succeeded in delivering. Our results suggest that one way of achieving such gains is to also give child-care providers the knowledge and skills to utilize available material and human resources in ways that are most productive for child development and well-being. The design of programs to facilitate this (and measurement of the degree to which this happens) continues to present big challenges globally and is likely to be a key hurdle in the path to achieving the SDG

on universal access to quality early childhood care by 2030. Our results, however, offer important promising evidence that even within the infrastructure available in LMIC contexts it is possible to implement scalable cost-effective designs of programs which boost the relevant dimensions of child-care quality and deliver significant improvements in child outcomes.

References

Agostinelli, F. and M. Wiswall (2016). Estimating the Technology of Children's Skill Formation. *Forthcoming JPE* (March).

Almond, D., J. Currie, and V. Duque (2018, dec). Childhood Circumstances and Adult Outcomes: Act II. *Journal of Economic Literature 56*(4), 1360–1446.

Andrew, A., O. Attanasio, R. Bernal, L. Cordona, S. Krutikova, D. M. Heredia, C. Medina, X. Peña, M. Rubio-Codina, and M. Vera-Hernandez (2018, jun). Evaluation of infant development centres: an early years intervention in Colombia. Technical Report June, International Initiative for Impact Evaluation, USA.

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

Araujo, M. C., P. Carneiro, Y. Cruz-Aguayo, and N. Schady (2016, aug). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics 131*(3), 1415–1453.

Araujo, M. C. and N. Schady (2015). Daycare Services: It's All about Quality. In S. Berlinski and N. Schady (Eds.), *The Early Years*, Chapter 4, pp. 91–121. New York: Palgrave Macmillan US.

Attanasio, O., H. Baker-Henningham, R. Bernal, C. Meghir, D. Pineda, and M. Rubio-Codina (2018, sep). Early Stimulation and Nutrition: The Impacts of a Scalable Intervention. Technical report, National Bureau of Economic Research, Cambridge, MA.

Attanasio, O., S. Cattan, E. Fitzsimons, C. Meghir, and M. Rubio-Codina (2015, feb). Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia. Technical report, National Bureau of Economic Research, Cambridge, MA.

Attanasio, O., C. Meghir, and E. Nix (2017). Human Capital Development and Parental Investment in India.

Barnett, W. S., J. T. Hustedt, A. H. Friedman, J. S. Boyd, and P. Ainsworth (2003). *The State of Preschool: 2003 State Preschool Yearbook*.

Barnett, W. S., J. T. Hustedt, A. H. Friedman, J. S. Boyd, and P. Ainsworth (2004). *The State of Preschool: 2004 State Preschool Yearbook*.

Behrman, J. R., J. Hoddinott, J. A. Maluccio, E. Soler-Hampejsek, E. L. Behrman, R. Martorell, M. Ramírez-Zea, and A. D. Stein (2014). What determines adult cognitive skills? Influences of pre-school, school, and post-school experiences in Guatemala. *Latin American Economic Review 23*(1).

Belfield, C., C. Farquharson, and L. Sibieta (2018, sep). 2018 Annual Report on Education Spending in England. Technical report, ESRC Centre for the Microeconomic Analysis of Public Policy (CPP) at IFS.

Berlinski, S., S. Galiani, and P. Gertler (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics 93*(1-2), 219–234.

Berlinski, S., S. Galiani, and M. Manacorda (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of Public Economics 92*(5-6), 1416–1440.

Bernal, R., O. Attanasio, X. Peña, and M. Vera-Hernández (2019). The effects of the transition from home-based childcare to childcare centers on children's health and development in Colombia. *Early Childhood Research Quarterly 47*, 418–431.

Bernal, R. and A. Camacho (2012). La política de primera infancia en el contexto de la equidad y movilidad social en Colombia. *Documentos Cede*, 2012–33.

Bernal, R. and C. Fernández (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*(C), 241–249.

Bernal, R., M. A. Martínez, and C. Quintero (2015). *Situación de niñas y niños Colombianos menores de cinco años, entre 2010 y 2013*, Volume 16.

Bernal, R. and S. M. Ramírez (2019). Improving the quality of early childhood care at scale: The effects of From Zero to Forever. *World Development 118*, 91–105.

Black, M. M., S. P. Walker, L. C. Fernald, C. T. Andersen, A. M. DiGirolamo, C. Lu, D. C. McCoy, G. Fink, Y. R. Shawar, J. Shiffman, A. E. Devercelli, Q. T. Wodon, E. Vargas-Barón, and S. Grantham-McGregor (2017). Early childhood development coming of age: science through the life course. *The Lancet 389*(10064), 77–90.

Blanden, J., K. Hansen, and S. Mcnally (2017). Quality in Early Years Settings and Children' s School Achievement. *CEP Discussion Paper No 1468*.

Britto, P. R., S. J. Lye, K. Proulx, A. K. Yousafzai, S. G. Matthews, T. Vaivada, R. Perez-Escamilla, N. Rao, P. Ip, L. C. H. Fernald, H. MacMillan, M. Hanson, T. D. Wachs, H. Yao, H. Yoshikawa, A. Cerezo, J. F. Leckman, and Z. A. Bhutta (2017, jan). Nurturing care: promoting early childhood development. *The Lancet 389*(10064), 91–102.

Britto, P. R., H. Yoshikawa, and K. Boller (2011). Quality of Early Childhood Development Programs in Global Contexts Rationale for Investment, Conceptual Framework and Implications for Equity. *25*(2).

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

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? Evidence from project star. *Quarterly Journal of Economics 126*(4), 1593–1660.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2018, dec). Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance. *Journal of Political Economy 126*(6), 2356–2409.

Culhane, J., F. Cunha, I. Elo, and Z. Pham (2016). Measuring Early Investments in Children Early.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006). Interpreting the Evidence on Life Cycle Skill Formation. Volume 1 of *Handbook of the Economics of Education*, pp. 697–812. Elsevier.

Cunha, F., J. J. Heckman, and S. M. Schennach (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica 78*(3), 883–931.

Danzer, V. A., M. F. Gerber, T. M. Lyons, and J. K. Voress (1991). *Daberon 2: Screening for School Readiness.*

Pro-Ed (Firm).

Datta Gupta, N. and M. Simonsen (2010). Non-cognitive child outcomes and universal high quality child care. *Journal of Public Economics 94*(1-2), 30–43.

Deming, D. (2009, jun). Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start. *American Economic Journal: Applied Economics 1*(3), 111–134.

Diamond, A. and C. Taylor (1996, may). 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), 315–334.

Dunn, L. M., E. R. Padilla, D. E. Lugo, and L. M. Dunn (1986). *Test de Vocabulario en Imágenes Peabody (TVIP)*. AGS Circle Pines, MN.

Engle, P. L., M. M. Black, J. R. Behrman, M. C. D. Mello, P. J. Gertler, L. Kapiriri, R. Martorell, and M. E. Young (2007). Child development in developing countries 3 Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. *The Lancet 369*(9557), 229–242.

Engle, P. L., L. C. Fernald, H. Alderman, J. Behrman, C. O'Gara, A. Yousafzai, M. C. de Mello, M. Hidrobo, N. Ulkuer, I. Ertem, and S. Iltus (2011, oct). Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. *The Lancet 378*(9799), 1339–1353.

Felfe, C. and R. Lalive (2018). Does early child care affect children's development? *Journal of Public Economics 159*(January), 33–53.

Garces, E., D. Thomas, and J. Currie (2002, aug). Longer-Term Effects of Head Start. *American Economic Review 92*(4), 999–1012.

Gerber, S. B., J. D. Finn, C. M. Achilles, and J. Boyd-Zaharias (2001, jun). Teacher Aides and Students' Academic Achievement. *Educational Evaluation and Policy Analysis 23*(2), 123–143.

Glewwe, P., E. Hanushek, S. Humpage, and R. Ravina (2011, oct). School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. Technical report, National Bureau of Economic Research, Cambridge, MA.

Hallam, R., B. Rous, S. Riley-Ayers, and D. Epstein (2011). *Teacher survey of early education quality*. New Brunswick, NJ: NIEER.

Hamadani, J. D., F. Tofail, A. Hilaly, S. N. Huda, P. Engle, and S. M. Grantham-McGregor (2010, mar). Use of Family Care Indicators and Their Relationship with Child Development in Bangladesh. *Journal of Health, Population and Nutrition 28*(1), 23–33.

Hanushek, E. A. (1999). Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis 21*(2), 143–163.

Hanushek, E. A. (2003, feb). The Failure of Inputbased Schooling Policies. *The Economic Journal 113*(485), F64–F98.

Hanushek, E. A. and S. G. Rivkin (2006). Chapter 18 Teacher Quality. In *Handbook of the Economics of Education*, Volume 2, pp. 1051–1078.

Hanushek, E. A. and S. G. Rivkin (2012, sep). The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics 4*(1), 131–157.

Harms, T., R. M. Clifford, and D. Cryer (1998). Early Childhood Environment Scale-Revised Edition.

Havnes, T. and M. Mogstad (2011, may). No Child Left Behind: Subsidized Child Care and Children's Long-Run Outcomes. *American Economic Journal: Economic Policy 3*(2), 97–129.

Havnes, T. and M. Mogstad (2015). Is universal child care leveling the playing field? *Journal of Public Economics 127*, 100–114.

Heckman, J., R. Pinto, and P. Savelyev (2013, oct). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review 103*(6), 2052–2086.

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

Heckman, J. J., S. H. Moon, R. Pinto, P. A. Savelyev, and A. Yavitz (2010, feb). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics 94*(1-2), 114–128.

Heckman, J. J. and S. Mosso (2014, aug). The Economics of Human Development and Social Mobility. *Annual Review of Economics 6*(1), 689–733.

Ichino, A., M. Fort, and G. Zanella (2019, apr). Cognitive and Non-Cognitive Costs of Daycare 0-2 for Children in Advantaged Families. *Journal of Political Economy*, 704075.

Jackson-Maldonado, D., V. A. Marchman, and L. C. Fernald (2013, jul). Short-form versions of the Spanish MacArthurBates Communicative Development Inventories. *Applied Psycholinguistics 34*(04), 837–868.

Jackson-Maldonado, D., D. J. Thal, L. Fenson, V. A. Marchman, T. Newton, B. T. Conboy, and L. Fenson (2003). *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual.* Baltimore: Brookes Publishing Co.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement 20*(1), 141–151.

Kline, P. and C. R. Walters (2016, nov). Evaluating Public Programs with Close Substitutes: The Case of Head Start. *The Quarterly Journal of Economics 131*(4), 1795–1848.

Kremer, M., C. Brannen, and R. Glennerster (2013). The Challenge of Education and Learning in the Developing World. *Science 340*(April), 297–301.

List, J., A. Shaikh, and Y. Xu (2016). Multiple Hypothesis Testing in Experimental Economics.

Mosteller, F. (1995). The Tennessee Study of Class Size in the Early School Grades. *The Future of Children 5*, 113–127.

Murnane, R. and A. Ganimian (2014). *Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Impact Evaluations.*

NICHD Early Child Care Research Network (2002, may). Child-Care Structure Process Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development. *Psychological Science 13*(3), 199–206.

Noboa-Hidalgo, G. E. and S. S. Urzúa (2012). The Effects of Participation in Public Child Care Centers: Evidence from Chile. *Journal of Human Capital 6*(1), 1–34.

Nores, M. and W. S. Barnett (2010, apr). Benefits of early childhood interventions across the world: (Under) Investing in the very young. *Economics of Education Review 29*(2), 271–282.

Nores, M., R. Bernal, and W. S. Barnett (2019, oct). Center-based care for infants and toddlers: The aeioTU randomized trial. *Economics of Education Review 72*, 30–43.

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

Economics 133(August 2017), 448–467.

Peisner-Feinberg, E. S., M. R. Burchinal, R. M. Clifford, M. L. Culkin, C. Howes, S. L. Kagan, and N. Yazejian (2001). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child development 72*(5), 1534–1553.

Peña, E. D. (2007, jul). Lost in Translation: Methodological Considerations in Cross-Cultural Research. *Child Development 78*(4), 1255–1264.

Rao, N., J. Sun, V. Pearson, E. Pearson, H. Liu, M. A. Constas, and P. L. Engle (2012). Is Something Better Than Nothing? An Evaluation of Early Childhood Programs in Cambodia. *Child Development 83*(3), 864–876.

Rao, N., J. Sun, J. M. S. Wong, B. Weekes, P. Ip, S. Shaeffer, M. Young, M. Bray, E. Chen, and D. Lee (2014). Early childhood development and cognitive development in developing countriess: A rigorous literature review. Technical report, Department for International Development.

Romano, J. and M. Wolf (2010). Balanced control of generalized error rates. *Annals of Statistics 38*(1), 598–633.

Romano, J. P. and M. Wolf (2005). Stepwise Multiple Testing as Formalized Data Snooping. *Econometrica 73*(4), 1237–1282.

Rosero, J. J. and H. Oosterbeek (2011). Trade-offs between Different Early Childhood Interventions: Evidence from Ecuador. Technical report.

Schrank, F. A., K. S. Mcgrew, M. L. Ruef, C. G. Alvarado, A. F. Muñoz-Sandoval, and R. W. Woodcock (2005). *Overview and technical supplement (Batería III Woodcock-Muñoz Assessment Service Bulletin No. 1)*. Itasca, IL: Riverside Publishing.

Skrondal, A. and S. Rabe-Hesketh (2009, jun). Prediction in multilevel generalized linear models. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(3), 659–687.

Squires, J., D. Bricker, and E. Twombly (2002). *Ages Stages Questionnaires: Social-emotional*. Brookes Pub. Co.

Squires, J., D. Bricker, and E. Twombly (2009). *Ages Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed, Child-Monitoring System*. Baltimore: Paul H. Brookes Publishing Co.

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

76–92.

The Economist (2019). Republicans and Democrats are taking early education more seriously.

Thomson, G. (1939). The factorial analysis of human ability. *British Journal of Educational Psychology 9*(2), 188–195.

Woodcock, R. W. (1977). Woodcock-Johnson Psycho-Educational Battery. Technical Report.

Yoshikawa, H., D. Leyva, C. E. Snow, E. Treviño, M. Clara Barata, C. Weiland, C. J. Gomez, L. Moreno, A. Rolla, N. D'Sa, and M. C. Arbour (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology 51*(3), 309–322.

Yoshikawa, H., C. Weiland, J. Brooks-Gunn, M. R. Burchinal, L. M. Espinosa, W. T. Gormley, J. Ludwig, K. A. Magnuson, D. Phillips, and M. J. Zaslow (2013). Investing in our future: The evidence base on preschool education.

Young, A. (2018, nov). Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*. *The Quarterly Journal of Economics 163*(7), F691–F703.

A Additional Tables

Table A.1: Predictors of Attrition at Endline

	(1)	(2)	(3)
HIM	-0.00629	-0.00350	-0.00279
	(0.0210)	(0.0207)	(0.161)
HIM+FE	-0.0184	-0.0228	-0.0899
	(0.0184)	(0.0180)	(0.142)
Male		0.0243*	0.0408
		(0.0140)	(0.0251)
Age at BL (months)		-0.00651***	-0.00792***
		(0.00152)	(0.00295)
Income (million COP)		0.0000127	0.0000212
		(0.00000821)	(0.0000140)
Maternal years of schooling		0.000217	0.000116
		(0.00266)	(0.00503)
ASQ BL factor		-0.00580	0.00391
		(0.00670)	(0.0106)
Language Development		0.00400	-0.00740
		(0.00733)	(0.0139)
ASQ SE factor BL		0.00821	0.00866
		(0.00714)	(0.0145)
=1 if Maternal Education missing		0.0248	0.0269
		(0.0692)	(0.0706)
=1 if ASQ factor missing		0.180	0.169
		(0.280)	(0.307)
HIM # Male			-0.0431
			(0.0343)
HIM # Age at BL (months)			0.00166
			(0.00416)
HIM # Income (million COP)			-0.00000256
			(0.0000221)
HIM # Maternal years of schooling			-0.00192
			(0.00644)
HIM # ASQ BL factor			-0.00807
			(0.0173)
HIM # Language Development			0.00917
			(0.0184)
HIM # ASQ SE factor BL			-0.0105
			(0.0191)
HIM+FE # Male			-0.00889
			(0.0340)
HIM+FE # Age at BL (months)			0.00242
			(0.00357)
HIM+FE # Income (million COP)			-0.0000215
			(0.0000174)
HIM+FE # Maternal years of schooling			0.00235
			(0.00707)
HIM+FE # ASQ BL factor			-0.0239*
			(0.0139)
HIM+FE # Language Development			0.0276
			(0.0170)
HIM+FE # ASQ SE factor BL			0.00762
			(0.0182)
Constant	0.0862***	0.246***	0.269**
	(0.0158)	(0.0637)	(0.116)
Observations	1987	1987	1987

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (clustered at HI level) in parentheses for regression of a dummy indicating attrition on: (1) treatment status, (2) treatment status and baseline characteristics and (3) treatment status fully interacted with baseline characteristics. Attrition defined as not having complete child development assessment data at endline.

Table A.2: Child Development Assessments

	Instrument used	Dimension of child development measured	Scored using IRT	In CLS factor	In pre-literacy skills factor	In SE skills factor
TVIP	Test de Vocabulario en Imgenes de Peabody	Receptive Language	Yes	Yes	Yes	
WM5	Woodcock-Muoz: Pruebas de Habilidades Cognitivas 5	Concept formation (cognition)	Yes	*		
WM12	Woodcock-Muoz: Pruebas de Habilidades Cognitivas 12	Fluid reasoning (cognition)		Yes		
WM14	Woodcock-Muoz: Pruebas de Aprovechamiento 14	Expressive language	Yes	Yes	Yes	
WM17	Woodcock-Muoz: Pruebas de Habilidades Cognitivas 17	Memory for words (cognition)	Yes	Yes	Yes	
WM21	Woodcock-Muoz: Pruebas de Aprovechamiento 21	Sound Awareness	Yes		*	
DAB	Daberon-II Screening for School Readiness	School readiness	Yes	Yes		
PTT	Pencil Tapping Task	Inhibitory control		Yes		
	ASQ:SE: Self Regulation	Self Regulation	Yes			Yes
	ASQ:SE: Compliance	Compliance	Yes			Yes
	ASQ:SE: Communication	Communication	Yes			Yes
	ASQ:SE: Adaptive Functioning	Adaptive Functioning	Yes			Yes
	ASQ:SE: Autonomy	Autonomy	Yes			Yes
	ASQ:SE: Affect	Affect	Yes			Yes
	ASQ:SE: Interaction with People	Interaction with People	Yes			Yes

Notes: *:initially included in factor (as set out in pre-analysis plan) but excluded due to low loading(¡0.4)

Table A.3: Validation of Child Cognitive, Language and School Readiness Development Assessments

		Correlation with									
			Baseline Child Development								
Instrument	Mean reliability	Age	Problem Solving	Communication	Gross Motor	Fine Motor	Socio-Individual	MacArthur-Bates	Wealth index	Mother's education	N
TVIP/Peabody	0.879	0.435***	0.193***	0.235***	0.089***	0.106***	0.112***	0.222***	0.241***	0.276***	1832
WM Cognitive 5	0.278	0.234***	0.050**	0.069***	-0.031	0.028	0.006	0.075***	0.03	0.079***	1839
WM Achievement 14	0.845	0.342***	0.147***	0.217***	0.076***	0.079***	0.055**	0.211***	0.222***	0.286***	1839
WM Cognitive 17	0.858	0.348***	0.103***	0.203***	0.085***	0.090***	0.083***	0.223***	0.077***	0.099***	1839
WM Achievement 21	0.395	0.117***	0.071**	0.090***	-0.036	-0.004	0.074**	0.088***	0.035	0.070**	1081
Daberon	0.925	0.517***	0.181***	0.297***	0.083***	0.128***	0.134***	0.240***	0.165***	0.230***	1839
WM Cognitive 12		0.364***	0.123***	0.199***	0.053**	0.060***	0.057**	0.242***	0.124***	0.199***	1839
Pencil Tapping Test		0.192***	0.090***	0.167***	0.014	0.071**	0.092***	0.180***	0.092***	0.129***	1081

Notes: Correlation coefficient significantly different from zero at: * p < 0.1, ** p < 0.05, *** p < 0.01. Mean reliability defined as $(1/se(\theta_m)^2)/(1+1/se(\theta_m)^2)$ where $se(\theta_m)$ refers to the standard error on the child's latent score computed through the empirical bayes' estimator (for measures scored using IRT only). For the correlation coefficient of the measures with age we use the IRT (or count) score before non-parametric standardization, for all other correlation coefficients we use the IRT (or count) score after non-parametric standardization (removing all age effects). Baseline measures standardised for age. Wealth index constructed using factor analysis of baseline asset ownership. Years of education of mother as measured at baseline.

Table A.4: Contemporaneous Correlation of Child Cognitive, Language and School Readiness Development Assessments

	TVIP/Peabody	WM Cognitive 5	WM Cognitive 12	WM Achievement 14	WM Cognitive 17	WM Achievement 21	Daberon	Pencil Tapping Test
TVIP/Peabody	1							
WM Cognitive 5	0.235***	1						
WM Cognitive 12	0.464***	0.151***	1					
WM Achievement 14	0.670***	0.228***	0.465***	1				
WM Cognitive 17	0.322***	0.0846***	0.283***	0.297***	1			
WM Achievement 21	0.215***	0.125***	0.128***	0.207***	0.206***	1		
Daberon	0.630***	0.254***	0.511***	0.590***	0.434***	0.292***	1	
Pencil Tapping Test	0.319***	0.169***	0.260***	0.271***	0.306***	0.137***	0.467***	1

Notes: Correlation coefficient significantly different from zero at: * p < 0.1, ** p < 0.05, *** p < 0.01. All measures are non-parametrically standardized to remove age effects.

Table A.5: Validation of Child Socio-Emotional Development Assessments

		Correlation with									
			Baseline Child Development								
	Mean reliability	Age	Self Regulation	Compliance	Communication	Adaptive Functioning	Autonomy	Affect	Interaction with People	Wealth index	Mother's education	N
Self Regulation	0.563	0.080***	0.228***	0.106***	0.107***	0.132***	0.012	0.122***	0.092***	0.132***	0.130***	1987
Compliance	0.522	0.017	0.136***	0.162***	0.082***	0.100***	0.008	0.054**	0.104***	0.072***	0.046*	1987
Communication	0.278	0.011	0.079***	0.048**	0.146***	0.098***	0.027	0.080***	0.094***	0.101***	0.141***	1987
Adaptive Functioning	0.206	0.078***	0.166***	0.060***	0.062***	0.131***	0.022	0.050**	0.080***	0.067***	0.092***	1987
Autonomy	0.289	-0.034	0.028	0.014	0.021	0.017	-0.024	0.052**	0.048**	0.094***	0.106***	1987
Affect	0.008	-0.002	0	0.032	-0.03	-0.044*	0.064***	-0.057**	0.001	-0.019	-0.085***	1987
Interaction with People	0.381	0.012	0.063***	0.015	0.083***	0.067***	-0.028	0.082***	0.112***	0.099***	0.074***	1987

Notes: Correlation coefficient significantly different from zero at: * p < 0.1, ** p < 0.05, *** p < 0.01. Mean reliability defined as $(1/se(\theta_m)^2)/(1+1/se(\theta_m)^2)$ where $se(\theta_m)$ refers to the standard error on the child's latent score computed through the empirical bayes' estimator (for measures scored using IRT only). For the correlation coefficient of the measures with age we use the IRT (or count) score before non-parametric standardization, for all other correlation coefficients we use the IRT (or count) score after non-parametric standardization (removing all age effects). Baseline measures standardised for age. Wealth index constructed using factor analysis of baseline asset ownership. Years of education of mother as measured at baseline.

Table A.6: Contemporaneous Correlation of Child Socio-Emotional Development Assessments

	Self Regulation	Compliance	Communication	Adaptive Functioning	Autonomy	Affect	Interaction with People
Self Regulation	1						
Compliance	0.372***	1					
Communication	0.346***	0.274***	1				
Adaptive Functioning	0.310***	0.178***	0.223***	1			
Autonomy	0.270***	0.187***	0.284***	0.183***	1		
Affect	-0.0697***	0.00487	0.00508	-0.0341	-0.00198	1	
Interaction with People	0.265***	0.238***	0.272***	0.185***	0.298***	-0.0444*	1

Notes: Correlation coefficient significantly different from zero at: * p < 0.1, ** p < 0.05, *** p < 0.01. All measures are non-parametrically standardized to remove age effects.

Table A.7: Factor Loadings for CLS and Pre-literacy Skills factors

	Factor Loadings		
	CLS	CLS (exc. PTT)	Pre-literacy Skills
TVIP	0.769	0.784	0.762
Daberon - total	0.805	0.783	
Pencil Tapping Test	0.472		
Woodcock Munoz - 12	0.587	0.614	
Woodcock Munoz - 14	0.73	0.753	0.748
Woodcock Munoz - 17	0.458	0.467	0.412
Eigenvalue	2.548	2.388	1.31

Notes: CLS refers to Cognitive, Language and School Readiness. Table presents factor loadings from three separate factor analyses. Allocation of measures to CLS and pre-literacy skills factors was pre-specified and set out in table A.2. CLS (exc. PTT) additionally constructed to obtain a CLS measure for the full sample since PTT only collected for children over 48 months of age. Factors with eigenvalue greater than one retained (Kaiser 1960); this led to one factor being retained for each analysis.

Table A.8: Factor Loadings for Socio-Emotional Skills

	Factor Loading
Self Regulation	0.612
Compliance	0.497
Communication	0.531
Adaptive Functioning	0.416
Autonomy	0.467
Interaction with People	0.481
Eigenvalue	1.526

Notes: Table presents factor loadings for factor analysis of socio-emotional skills measures. Allocation of measures to socio-emotional skills factors was pre-specified and set out in table A.2. Factors with eigenvalue greater than one retained (Kaiser 1960); this led to one factor being retained.

Table A.9: Factor Loadings for Play Materials

	Factor Loading
Musical toys	0.485
Building toys	0.664
Painting/writing toys	0.3
Movement toys	0.334
Fantasy toys	0.3
Story books	0.518
Colouring books	0.505
Shape toys	0.687
Eigenvalue	1.963

Notes: Table presents factor loadings for factor analysis of play materials measures. Factors with eigenvalue greater than one retained (Kaiser 1960); this led to one factor being retained.

Table A.10: Factor Loadings for Play Activities with Mother

	Factor Loading
Activities with books (minutes)	0.54
Telling stories (minutes)	0.526
Play with child's toys (minutes)	0.465
Draw, paint etc (minutes)	0.33
Singing (times)	0.456
Go to park, sqaure etc (times)	0.332
Go out to do chores (times)	0.167
Go out to play (times)	0.372
Naming, counting etc (times)	0.449
Talking about day etc (times)	0.296
Watching TV (times)	0.241
Dancing (times)	0.381
Eigenvalue	1.87

Notes: Table presents factor loadings for factor analysis of play activities with mother measures. Factors with eigenvalue greater than one retained (Kaiser 1960); this led to one factor being retained.

Table A.11: Factor Loadings for Play Activities with Father

	Factor Loading
Activities with books (minutes)	0.653
Telling stories (minutes)	0.634
Play with child's toys (minutes)	0.571
Draw, paint etc (minutes)	0.537
Singing (times)	0.536
Go to park, sqaure etc (times)	0.437
Go out to do chores (times)	0.216
Go out to play (times)	0.457
Naming, counting etc (times)	0.529
Talking about day etc (times)	0.42
Watching TV (times)	0.449
Dancing (times)	0.404
Eigenvalue	2.997

Notes: Table presents factor loadings for factor analysis of play activities with father measures. Factors with eigenvalue greater than one retained (Kaiser 1960); this led to one factor being retained.

Table A.12: Impact of HIM and HIM+FE on all Directly Assessed Measures of Child Development

	Receptive Language (TVIP)	Fluid reasoning (WM12)	Expressive language (WM14)	Memory for words (WM17)	School Readiness (DAB)	Inhibitory Control (PTT)
	(1)	(2)	(3)	(4)	(5)	(6)
HIM only	-0.084	0.006	-0.084	-0.072	-0.003	0.038
	(0.062)	(0.056)	(0.069)	(0.057)	(0.066)	(0.072)
	$[p = 0.169, p^{FW} = 0.540]$	$[p = 0.916, p^{FW} = 0.991]$	$[p = 0.221, p^{FW} = 0.545]$	$[p = 0.203, p^{FW} = 0.580]$	$[p = 0.968, p^{FW} = 0.968]$	$[p = 0.588, p^{FW} = 0.909]$
	$\{p = 0.201\}$	$\{p = 0.913\}$	$\{p = 0.215\}$	$\{p = 0.255\}$	$\{p = 0.966\}$	$\{p = 0.599\}$
HIM+FE	-0.011	0.148***	0.099	0.021	0.055	0.100
	(0.063)	(0.057)	(0.065)	(0.053)	(0.062)	(0.072)
	$[p = 0.853, p^{FW} = 0.853]$	$[p = 0.006, p^{FW} = 0.030]$	$[p = 0.126, p^{FW} = 0.401]$	$[p = 0.681, p^{FW} = 0.899]$	$[p = 0.372, p^{FW} = 0.688]$	$[p = 0.165, p^{FW} = 0.452]$
	$\{p = 0.863\}$	$\{p = 0.011\}$	$\{p = 0.097\}$	$\{p = 0.737\}$	$\{p = 0.356\}$	$\{p = 0.180\}$
HIM + FE vs. HIM	0.073	0.142**	0.183***	0.093*	0.058	0.061
	(0.063)	(0.070)	(0.064)	(0.056)	(0.064)	(0.073)
	$[p = 0.246, p^{FW} = 0.513]$	$[p = 0.046, p^{FW} = 0.171]$	$[p = 0.005, p^{FW} = 0.023]$	$[p = 0.097, p^{FW} = 0.287]$	$[p = 0.362, p^{FW} = 0.583]$	$[p = 0.410, p^{FW} = 0.410]$
	$\{p = 0.245\}$	$\{p = 0.036\}$	$\{p = 0.001\}$	$\{p = 0.137\}$	$\{p = 0.353\}$	$\{p = 0.391\}$
N	1819	1825	1825	1825	1825	1073

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) $\{randomization - t\}$. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across all outcomes in table. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, tester effects and baseline scores for MacArthur Bates CDI and each sub-scale of the ASQ-III. All measures scaled to have a mean of zero and standard deviation of one in the control group. All measures are age-standardized. See Table A.2 for details of measures.

Table A.13: Proportion of Teachers Completing FE Teacher Training by BL Number of Children

	Self-Reported Data	Administrative Data
	(1)	(2)
Bigger BL HI	0.280	0.255
Smaller BL HI	0.452	0.507
p-value on difference	0.037	0.000

Notes: Table shows number of teachers who completed the FE training relative to the number of teachers in the HI at baseline for HIs allocated to the HI+FE group. Self reported data comes from teacher questionnaires at endline. Administrative data comes for records kept by the Colombian National University.

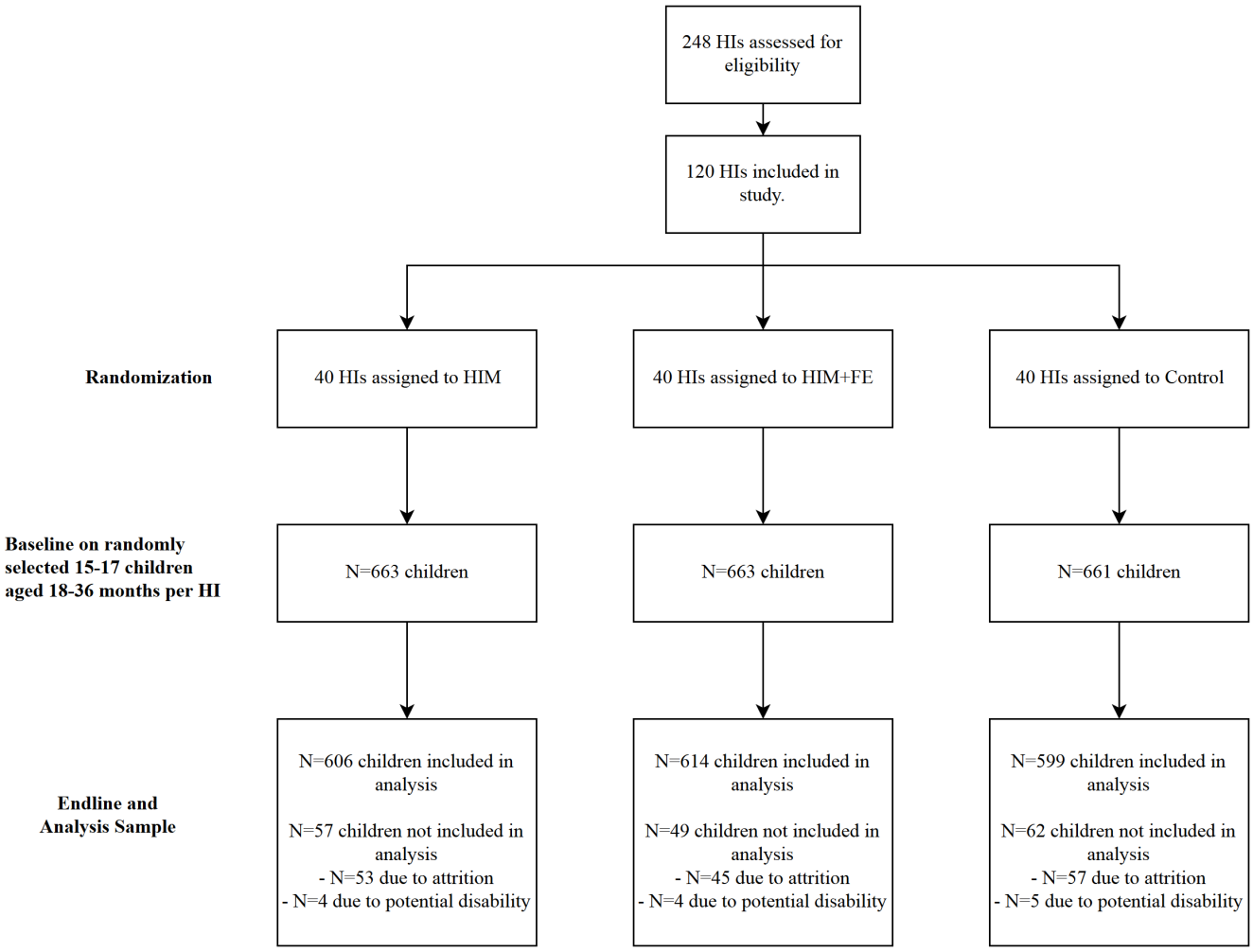
Table A.14: Heterogeneity by HI Characteristics

	Heterogeneity by HI Size	
	CLS	Pre-literacy
	(1)	(2)
HIM only	0.092	0.026
	-0.099	-0.097
	$[p = 0.352, p^{FW} = 0.508]$	$[p = 0.800, p^{FW} = 0.800]$
	$\{p = 0.354\}$	$\{p = 0.791\}$
HIM only X Bigger BL HI	-0.283*	-0.266*
	-0.16	-0.158
	$[p = 0.076, p^{FW} = 0.192]$	$[p = 0.088, p^{FW} = 0.183]$
	$\{p = 0.103\}$	$\{p = 0.095\}$
HIM+FE	0.253**	0.186*
	-0.114	-0.113
	$[p = 0.029, p^{FW} = 0.069]$	$[p = 0.100, p^{FW} = 0.165]$
	$\{p = 0.023\}$	$\{p = 0.073\}$
HIM+FE X Bigger BL HI	-0.244	-0.295*
	-0.17	-0.154
	$[p = 0.152, p^{FW} = 0.152]$	$[p = 0.054, p^{FW} = 0.111]$
	$\{p = 0.153\}$	$\{p = 0.050\}$
HIM+FE vs. HIM	0.161	0.161
	(0.106)	(0.101)
	$[p = 0.132, p^{FW} = 0.257]$	$[p = 0.112, p^{FW} = 0.261]$
	$\{p = 0.169\}$	$\{p = 0.130\}$
HIM+FE vs. HIM X Bigger BL HI	0.039	-0.029
	(0.168)	(0.156)
	$[p = 0.812, p^{FW} = 0.948]$	$[p = 0.847, p^{FW} = 0.847]$
	$\{p = 0.831\}$	$\{p = 0.853\}$
N	1071	1819

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01 using bootstrap p-values. Standard errors (bootstrapped) in parentheses. Single-hypothesis two-sided p-values calculated 2 ways: (i) [*bootstrap*] and (ii) {$randomization-t$}. Adjusted two-sided p-values (p^{FW}) are equivalent to bootstrap p-values but adjusted for testing each null hypothesis (null impact of HIM, HIM+FE and the comparison) on multiple outcomes through the stepwise procedure described in List, Shaikh, and Xu (2016). In particular, we correct for multiple testing across our two pre-specified child development outcomes and, for each of the two heterogeneity analyses separately, across the two subgroups. Bootstrapped statistics use block-bootstraps, resampling triplets with replacement, and comprise 5000 iterations. Randomization inference (5000 iterations) accounts for clustering at HI level and stratification by triplets. All estimates control for gender, city effects, tester effects and baseline scores for MacArthur Bates CDI 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 4. Age effects removed from standardized scores prior to factor construction. Each factor is constructed using the following standardized scores: (i) Cog, Lang and Sch (all measures): WM12, WM14, WM17, TVIP, DAB and PTT, (ii) Cog, Lang and Sch (exc. PTT): WM12, WM17, TVIP, WM14 and DAB, and (iii) Pre-literacy skills: WM14, WM17 and TVIP. See Table A.2 for details of measures and factors. Bigger BL HI implies child was in an HI with above median (155) number of children at baseline.

B Additional Figures

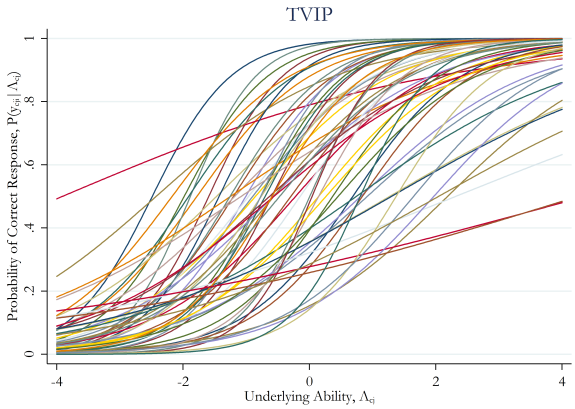
Figure B.1: Study Flow Diagram



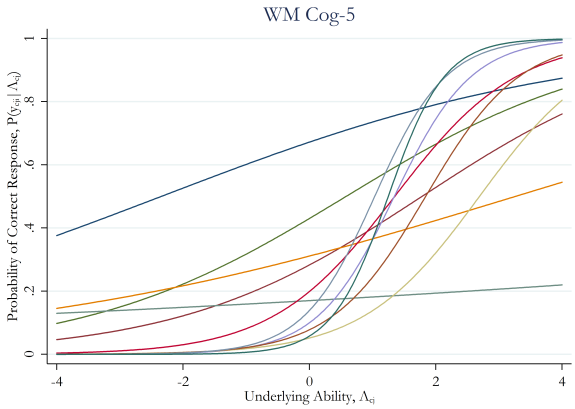
Notes: Wherever there were 15, 16 or 17 18-36 month olds enrolled in an HI at baseline, we included them all in the study. Whenever there were 18 or more we randomly selected 17 for inclusion. As pre-specified, we dropped children with a 'potential disability' meaning that their Z-score was ≤ -3 for at least one child development measure.

Figure B.2: Item Characteristic Curves (ICCs) for Cognitive, Language and School Readiness Assessments

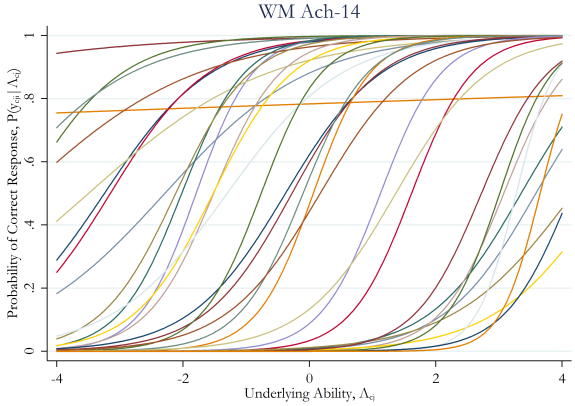
(a) Receptive Language



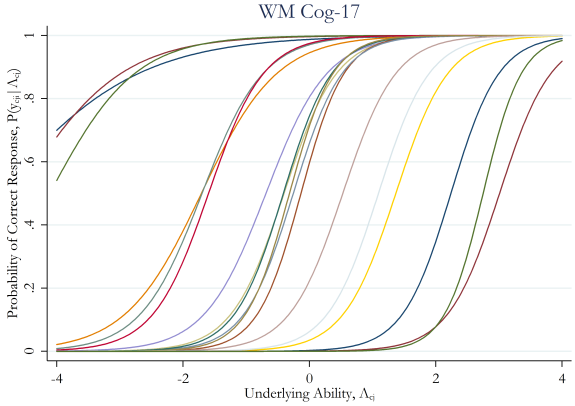
(b) Concept Formation



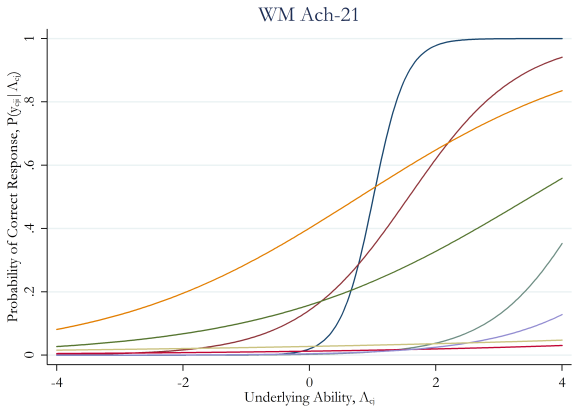
(c) Expressive Language



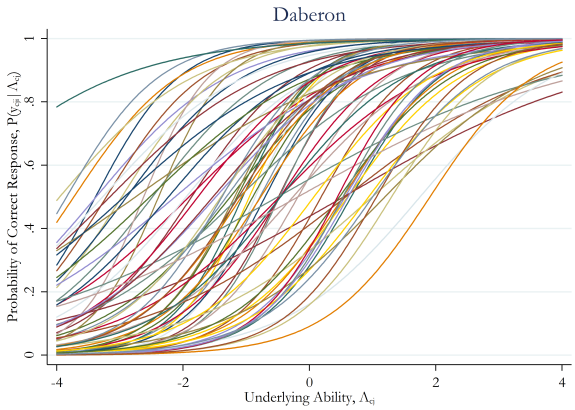
(d) Memory for Words



(e) Sound Awareness

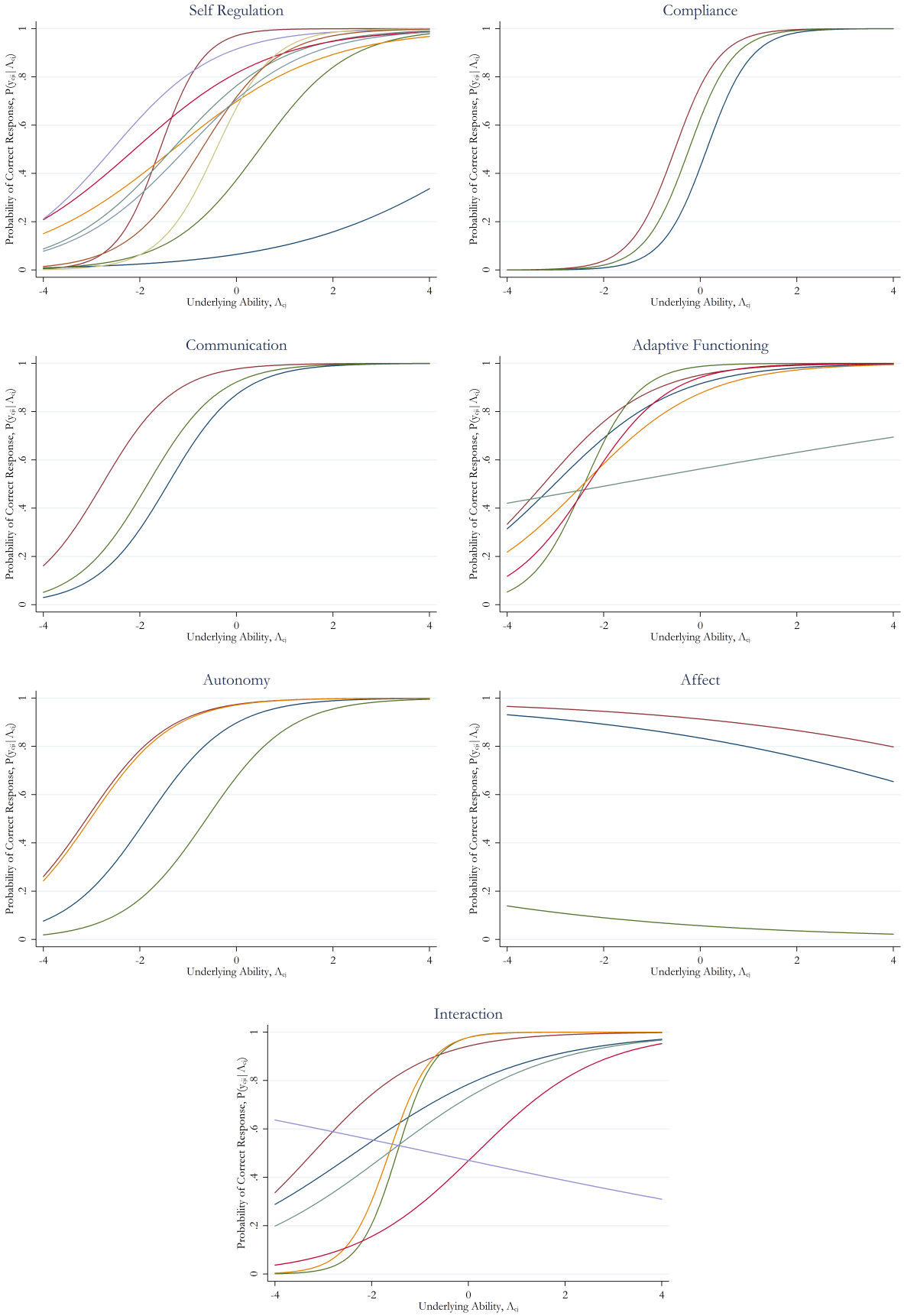


(f) School Readiness



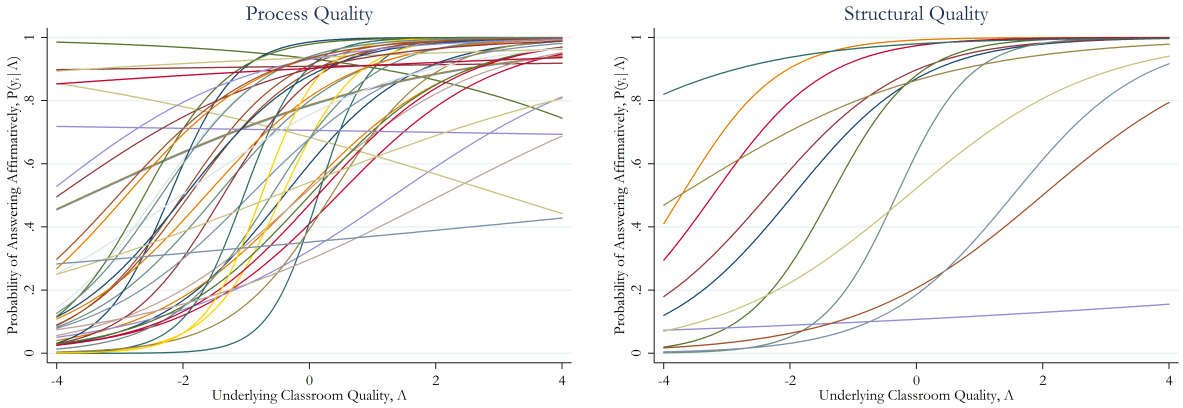
Notes: Figures show Item Characteristic Curves (ICCs) for each directly-assessed measure of child development with item-level data. Using results from our 2-parameter IRT model, they plot, for each item i, the probability that a child answers correctly as a function of child c's underlying ability Λ_{cj} in skill j. This probability is: $P(y_{cji}|\Lambda_{cj}) = exp(\alpha_i + \beta_i\Lambda_{cj})/(1 + exp(\alpha_i + \beta_i\Lambda_{cj}))$ See Appendix for details of IRT procedure. Details of measures in Table A.2.

Figure B.3: Item Characteristic Curves (ICCs) for Socio-Emotional Development



Notes: Figures show Item Characteristic Curves (ICCs) for each maternal-report measure of socio-emotional development. Using results from our 2-parameter IRT model, they plot, for each item i, the probability that a child's mother answers affirmatively as a function of child c's underlying ability Λ_{cj} in skill j. This probability is: $P(y_{cji}|\Lambda_{cj}) = exp(\alpha_i + \beta_i\Lambda_{cj})/(1 + exp(\alpha_i + \beta_i\Lambda_{cj}))$ See Appendix for details of IRT procedure. Details of measures in Table A.2.

Figure B.4: Item Characteristic Curves (ICCs) for Classroom Quality Assessments (ECERS)



Notes: Figures show Item Characteristic Curves (ICCs) for our two dimensions of preschool quality, as measured by the ECERS-R. Using results from our 2-parameter IRT model, they plot, for each item i, the probability that the item was answered affirmatively as a function of the underlying preschool quality Λ_{cj} in domain j. This probability is: $P(y_{cji}|\Lambda_{cj}) = exp(\alpha_i + \beta_i\Lambda_{cj})/(1 + exp(\alpha_i + \beta_i\Lambda_{cj}))$ See Appendix for details of IRT procedure. Details of ECERS-R measure construction in Appendix D.3.

C Hogares Infantiles in the Context of ECE Policy in Colombia

Hogares Infantiles (HIs) is the oldest, and until 2010 the only, public center-based early education program in Colombia. HIs are partly-subsidized public preschools which are run by the government's National Family Welfare Agency, *Instituto Colombiano de Bienestar Familiar* (ICBF). There are 1,008 HIs across Colombia which have enrolled 125,000 children per year over the past decade.

HIs are typically located in fairly well-equipped community centers. They typically have between 3 and 10 teachers each and each teacher cares for a class of, on average, 30 children. HIs provide care for 8 hours a day, 5 days a week. In addition to providing early education, they also provide children with up to 60% of daily nutritional requirements. The out of pocket costs to parents or carers is around USD10-25 per month.

HIs serve children aged 6 months to 5 years of age. However, the HIs only take children younger than two when "it is proven that they do not have a responsible adult to care for them". Correspondingly, the vast majority of children in HIs are between 2 and 5 years. HIs are intended to target children "from vulnerable working families". Priority is given to children belonging to households whose SISBEN score (Colombias proxy means test for allocating social welfare programs) is puts them below the poverty line, and to children who are at the highest risk of not being able to attend childcare services and being left alone at home.

In 2011, the government launched the national early childhood strategy From Zero to Forever (De Cero a Siempre, DCAS henceforth) aimed at increasing access and improving the quality of early childhood services provided to poor children. The objective was to deliver high-quality early childhood services for 1.2 million children under the

age of 6 in poverty. As part of the DCAS strategy, the government of Colombia implemented a quality upgrade of the HI program starting in 2013.

D Measurements

In this appendix we provide further details of our construction of both measures of child development and measures of preschool quality.

D.1 IRT Scoring of Child Development Assessments

As discussed in Section 3.1, we utilize item-level data on how each child (or mother for socio-emotional measures) answered every question to predict the underlying latent constructs that each instrument captured. We use a two-parameter item response theory (IRT) model to estimate the latent construct or skill that determines how a child performed on the items comprising that instrument. Specifically, let whether or not child c answers item i correctly in instrument j depend on the child c's underlying latent skill j. Call this latent skill Λ_c^j and assume that it is normally distributed across children with zero mean and unit variance. Allowing different items to differ in both their difficulty – how likely is a child of average skill to answer correctly – and their discriminatory power – how sensitive the probability of answering correctly is to a child's skill – we model the process determining children's answers as:

$$y_{ci}^j = \begin{cases} 1 & if \quad \alpha_i + \beta_i \Lambda_c^j + \epsilon_{ci}^j > 0 \\ 0 & otherwise \end{cases} \tag{D.1}$$

where $y_{ci}^j = 1$ if child c answers item i correctly in instrument j and $y_{ci}^j = 0$ otherwise. We assume that the errors ϵ_{ci}^j are independently and identically distributed as type-I extreme value with variance $\pi^2/6$ across items and individuals. In this set-up α_i represents an item's difficulty – the higher is α_i the easier an item is – and β_i represents its discriminatory power – the rate at which the probability of answering correctly changes with a child's skill. We estimate $\alpha_i, \beta_i, \Lambda_c^j$ for skill j in two stages. First, we estimate α_i, β_i for $i \in \{1, 2, ...I\}$ using maximum likelihood, integrating out the unobserved Λ_c^j using adaptive Gauss–Hermite quadrature. Second, we use empirical Bayes estimators of the latent skill Λ_c^j of each child c (Skrondal and Rabe-Hesketh 2009). We take the mean of the

empirical posterior distribution of Λ_c^j conditional on the child's item responses $\boldsymbol{y}_c^j$, imposing our estimates $\hat{\alpha}_i, \hat{\beta}_i$. Again, we use adaptive Gauss–Hermite quadrature to integrate over the prior distribution of Λ_c^j. [37]

D.2 Age standardization

We age-standardize our child-development scores $\hat{\Lambda}_c^j$. To do so, we non-parametrically estimate the age-specific mean, $\hat{E}(\hat{\Lambda}_c^j | age_c, T_c^{control} = 1))$, and variance, $\hat{V}(\hat{\Lambda}_c^j | age_c, T_c^{control} = 1))$, of the control group ($T_c^{control} = 1)$) using locally weighted regressions of $\hat{\Lambda}_c^j$ and $(\hat{\Lambda}_c^j - \hat{E}(\hat{\Lambda}_c^j | age_c, T_c^{control} = 1)))^2$ on age. We then construct age standardized z-scores as follows:

$$m_c^j = \frac{\hat{\Lambda}_c^j - E(\hat{\Lambda}_c^j | age_c, T_c^{control} = 1)}{\sqrt{\hat{V}(\hat{\Lambda}_c^j | age_c, T_c^{control} = 1)}} \tag{D.2}$$

D.3 Constructing Preschool Quality Measures from the ECERS-R

As discussed in Section 3.2, the ECERS-R is comprised of 43 individual items which each measuring a different aspect of quality, for example encouraging children to communicate. Each item is formed of around 10 sub-items grouped under the headings inadequate, minimal, good and excellent to which the observer must answer true or false. However, the official procedure, which we followed, for administering the sub-items turned out to be poorly suited to our context. In particular, the procedure dictates that if the observer answers true to any of the sub-items listed under the inadequate heading, all of which are negatively phrased, then on that item the classroom is categorized as inadequate and the observer does not proceed to answer the sub-items in the other categories. Similar stopping rules apply to the other categories. The difficulty lies in the fact that often a positive answer to a negatively-phrased inadequate item would not seem to preclude a positive answer to a positively-phrased minimal, good or excellent item. For example, in the previously mentioned item, answering true to very few materials accessible that encourage children to communicate does not preclude that the observer could have also answered positively to staff balance listening and talking appropriately for age and abilities of children during communication activities but following the official procedure if an observer did answer positively to the former then they would not answer the latter at all. Given that for many, or even most, of the classrooms in our sample many of the inadequate items were judged to be true we have a huge amount of missing observations for later sub-items. Indeed, averaged across all classrooms 41.9% of the 129 minimal sub-items and 78.9% of the 136 good sub-items are not answered. This creates unstable

[37] To ensure parameter estimates were stable we only used items attempted by at least 20% of the sample.

IRT parameter estimates when including sub-items with many missing values in a relatively small sample.

Given the number of non-random missing values for items categorized as minimal, good and excellent we drop these items from our analysis and construct measures that use only the inadequate items that are, by construction, non-missing for all classrooms. This unfortunately implies that the sub-items making up our quality measures are the absence of poor practices rather than the presence of good practices. We drop items with very low variance (≤ 0.10) and estimate separate underlying latent constructs for a classrooms process quality and structural quality using a two-parameter IRT model, as described in equation (D.1). In doing so, we group together all sub-items in the Personal Care Routines, Language and Literacy, Learning Activities, Interaction and Program Structure subscales to construct a measure of process quality while using items in the Space and Furnishings subscale to construct a measure of structural quality. To increase the sample size for estimating IRT parameters we pool ECERS-R measures from baseline and endline giving a total sample of 296 observations. We drop items that were found to be preventing model convergence.

E Costs of Pedagogical Training Program

We argue that the pedagogical training program is the key component of FE in terms of generating impacts on children's cognition, language and school readiness. FE provided us with the total cost of this component (COP 419546284, or USD 233081 at February 2013 exchange rate of 1800 COP/USD). With this budget, FE provided completed training for 99 teachers from the 40 HIs in the HIM+FE treatment arm.[38] This represented 32% of teachers who worked in those 40 HIs. We thus estimate that the initial one-off cost to roll out the pedagogical training program to new HIs, at the same intensity as achieved the impacts we see in this study (i.e. training 32% of teachers), would be USD 5827 per HI or USD 35 per child attending an HI.

However, it is unreasonable to assume that the same intensity of training program would be required year after year for FE to sustain its impacts on successive cohorts. Rather, we calculate the costs of maintaining a ratio of training 32% of staff, which implies providing training for 32% of new staff, and the costs of providing a yearly refresher training to all teachers who have already been trained which we assume would cost 25% of the costs of the full training. Given these assumptions, we estimate that the ongoing cost per center of maintaining the results

[38] More teachers began training however, for consistency, we calculate costs relative to the number completing. Presuming the drop-out rates seen during the study are similar to what they would be if the program were scaled, this makes no difference. We also note that in some cases other staff (headteachers, teaching assistants etc) also completed the training. To be as conservative as possible in calculating costs we simply calculate cost per teacher completing rather than cost per person. This means that our projected costs also allow the same proportion of other staff to receive training as they did in the trial.

of the pedagogical training program would be USD 2206 per year and the cost per child would be USD 13 per year. All data, assumptions and formulae used in these calculations are shown in Table D.1.

In interpreting these costs, there are two points to note. First, to train 100% of teachers, rather than 32%, would be more costly. However, since the benefits found in this study were from training 32% of teachers we consider this the most meaningful cost. We would expect benefits to children's development to be larger if a greater proportion of teachers were trained. Second, our cost figures are based on 32% of *all* teachers in the center receiving the training, irrespective of the age they teach, and the cost per child figure is based on the total number of children in the center. Our study children were between 1 and 3 at baseline and between 3 and 5 at endline. Given only 2.7% of teachers report that they primarily teach children younger than one year, we consider the training of all teachers relevant for generating the treatment effect. Moreover, we note that teachers' propensity to complete the training appears independent of the age of the children they teach. Therefore, we include all teachers and children of all ages in the costing.

Table D.1: Rough costs of scaling Pedagogical Training component of FE

			Source
	Costs of pedagogical training program		
(1)	Total cost of FE pedagogical training program	USD 233,081	FE
(2)	Number of teachers who completed training	99	FE
(3)	Cost per teacher completing training	USD 2,354.35	(1)/(2)
	Actual intensity of pedagogical training program in FE treatment arm		
(4)	Total number of teachers in HIs allocated to HIM+FE treatment arm	313	Endline survey
(5)	Proportion of teachers who completed training in HIM+FE treatment arm	0.32	(2)/(4)
	Projected one-off cost per center of training 32% of teachers		
(6)	Average number of children per center	166	Endline survey
(7)	***One-off cost per center of training 32% of teachers***	***USD 5,827.03***	***(1)/40***
(8)	***One-off cost per child of training 32% of teachers***	***USD 35.10***	***(7)/(6)***
	Projected ongoing cost of maintaining 32% of teachers trained (inc. yearly refresher training)		
(9)	Proportion of one-off costs required to complete yearly refresher training	0.25	Assumption
(10)	Average number of teachers per HI	7.05	Endline survey
(11)	Average number of new teachers per HI per year (number who joined in 2014)	1.6	Endline survey
(12)	Yearly cost per center of training 32% of new teachers	USD 1,191.47	(3)x(5)x(11)
(13)	Yearly cost per child of training 32% of new teachers	USD 7.18	(12)/(6)
(14)	Yearly cost per center of refresher training for all previously trained teachers	USD 1,014.61	[(10)-(11)]x(9)x(3)x(5)
(15)	Yearly cost per child of refresher training for all previously trained teachers	USD 6.11	(14)/(6)
(16)	***Yearly cost per center of maintaining 32% of teachers trained (inc. yearly refresher training)***	***USD 2,206.08***	***(12)+(14)***
(17)	***Yearly cost per child of maintaining 32% of teachers trained (inc. yearly refresher training)***	***USD 13.29***	***(13)+(15)***