

Preschool Quality and Child Development

Alison Andrew

Oxford University and Institute for Fiscal Studies

Orazio P. Attanasio

Yale University and National Bureau of Economic Research

Raquel Bernal

Universidad de los Andes

Lina Cardona Sosa

World Bank

Sonya Krutikova

Manchester University and Institute for Fiscal Studies

Marta Rubio-Codina

Inter-American Development Bank and Institute for Fiscal Studies

Globally, preschool enrollment has surged, but its quality is often poor. We evaluate strategies to improve quality of public preschools in Colombia. The first, designed by the government and rolled out nationwide,

We thank Diana Pérez Lopéz and Diana Martínez Heredia for excellent research assistance and gratefully acknowledge the contributions of Carlos Medina and Marcos Vera-Hernández to

Electronically published May 31, 2024

Journal of Political Economy, volume 132, number 7, July 2024.

© 2024 The University of Chicago. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial reuse of the work with attribution. For commercial use, contact journalpermissions@press.uchicago.edu. Published by The University of Chicago Press.
<https://doi.org/10.1086/728744>

provided extra funding, mainly earmarked for hiring teaching assistants. The second also offered low-cost training for existing teachers. The first intervention had no effect on child development, while the second improved children's cognitive development, especially for more disadvantaged children. This pattern can be explained by the interventions affecting teachers' behavior differently. The first led teachers to reduce their classroom time, including learning activities, while additional training offset the adverse effect on learning activities and improved teaching quality.

I. Introduction

It is now widely accepted that well-designed early childhood education (ECE) programs can have substantial and long-lasting positive effects on children (Elango et al. 2015). Consequently, there is significant momentum behind investing in early years education in both lower- and higher-income countries. Universal access to quality early childhood care by 2030 is one of the Sustainable Development Goals, and, globally, enrollment in preprimary education is rising fast. Enrollment increased from 29% in 1990 to 49% in 2015.¹ However, as governments expand coverage of ECE programs, quality should be a first-order concern. If not of good quality, these programs may deliver few benefits for child development and can even be inferior to home care (Britto, Yoshikawa, and Boller 2011; Engle et al. 2011; Rosero and Oosterbeek 2011; Araujo and Schady 2015; Fort, Ichino, and Zanella 2020).

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. We thank our editor James Heckman and three anonymous reviewers for extremely useful comments and suggestions. This research was funded by the International Initiative for Impact Evaluation (3rd) and Fundación Éxito. Attanasio acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 695300-HKADeC-ERC-2015-AdG). Andrew and Krutikova acknowledge funding from the Economic and Social Research Council Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. Bernal acknowledges funding from the British Academy Visiting Fellowship (VF1 10124). We thank the Jacobs Foundation for hosting us at the Marbach Residence Program in 2017 where we made significant progress on this project. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing up results. The views here presented do not represent the Inter-American Development Bank, The World Bank, their boards of directors, or the countries they represent. Ethics Committees at Universidad de los Andes and University College London approved the study's protocol in 2013. This paper was edited by James J. Heckman.

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 preprimary education, regardless of age, as a percentage of preprimary-age population. It classifies preprimary 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."

This issue is particularly relevant for lower- and middle-income countries (LMICs), where, according to the limited available evidence, ECE services are of widely variable quality, with many children receiving poor-quality center-based care (Araujo and Schady 2015; Yoshikawa et al. 2018). Many LMICs are resorting to adding preprimary classes to existing primary schools without allocating sufficient resources or expertise to ensure that the provision of high-quality education tailored to the needs of young children (Neuman and Okeng'o 2019). The risk is that the ongoing scale-ups of ECE provision will replicate the problems of low learning levels observed in the aftermath of primary- and secondary-education expansions in LMICs if they achieve high enrollment into poor-quality programs (Pritchett 2013; Glewwe and Muralidharan 2016; World Bank 2018; Singh 2020). Therefore, there is a need to design interventions that enhance the quality of existing ECE services. However, evidence on how to do this in a cost-effective way is scant, especially in LMICs. Most of the existing research focuses on estimating the overall impact of ECE programs relative to home care; few studies try to understand which aspects of ECE programs are most important for child development or how effective specific improvements to existing programs are. The evidence that we do have (mainly for the United States) suggests that not all commonly adopted approaches yield the expected benefits (Joo et al. 2020).

Our study adds to this evidence. We worked with the government of Colombia to evaluate the impact of two interventions designed to improve the quality of public preschools attended by relatively disadvantaged children. We provide evidence on the impacts of the interventions on child development. We then explore potential mechanisms, using data on the quality of the classroom learning environment and time use of teachers. We ground this analysis in a discussion of different margins on which teachers might respond to such interventions, as well as the effects that economic theory would predict.

The first of the two interventions, which we label “HIM” (Hogares Infantiles Mejorados) in line with the acronym that the government used for it, was designed by the Colombian government and rolled out nationwide. It provided preschools with additional funds, which were primarily earmarked for hiring teaching assistants (TAs). The second intervention complemented the first by additionally providing professional development training for existing preschool teachers. We label the additional training component “FE” in recognition of the contribution of Fundación Éxito, the Colombian nongovernmental organization who developed and ran the training program in partnership with the Colombian National University. We label the second combined intervention “HIM+FE”.

We find that HIM had no positive impacts on child development, despite high compliance and the fact that it represented a large increase in government investment in preschools. However, we show that, at moderate extra

cost, HIM+FE did have significant positive impacts on child development. After 18 months of exposure to the HIM+FE program, we find an improvement in children's cognitive development relative to the control group equivalent to 0.16 of the control group standard deviation (SD); relative to the HIM-only arm, the addition of FE improved child development by 0.17 SD. In line with several other studies (Havnes and Mogstad 2015; Cornelissen et al. 2018; Felfe and Lalive 2018), we find that children from poorer families benefited the most; these children's cognitive development improved by, on average, nearly 0.30 SD.

In addition to the impacts on children's development, we study the effects that the two interventions had on how teachers allocated their time to different activities, both within the classroom and outside of it. At baseline, the average teacher worked more than their contracted hours and had significant administrative duties. It is therefore plausible that teachers could have responded to the interventions by adjusting their total classroom time either upward or downward. Teachers also have a high degree of autonomy over how they manage their class and how they split their time among different teaching activities, as well as over what they instruct the TAs to do. Therefore, we explore how the two interventions affect the mix of activities that teachers and TAs perform, focusing on the distinction between learning and care activities.

Using novel data that capture teachers' day-to-day activities, we find that teachers responded to the HIM program by reducing their overall involvement in classroom activities. We observe that they reduced their hours of overtime. Moreover, they reduced their involvement not only in care activities but also in learning-focused activities, which we show are highly correlated with children's development. We suggest that this scaling back of their own efforts in response to being given TAs would be expected if teachers value highly a marginal reduction in their overtime and would be particularly large if teachers perceive that TAs are highly substitutable with themselves. Moreover, the fact that we see teachers reducing their learning activities, in addition to their caring ones, could suggest that they are not fully exploiting their presumed comparative advantage, relative to their TAs, in learning activities. The addition of FE, however, induced teachers to increase the time that they allocated to the job, increased their involvement in learning activities and improved the quality of teaching as directly observed by trained psychologists. This response suggests that the addition of FE increased how useful teachers thought that the marginal time they spent on learning activities was for child development relative to that devoted to care activities. This may reflect a shift in real productivity of the teachers in implementing learning activities or an upward revision in their perception of the importance for child development of devoting time to learning activities (Caucutt, Lochner, and Park 2017; Cunha, Elo, and Culhane 2022).

Taken together, the impacts we find on child development and teacher time allocation, suggest that, given teachers' preferences and their perceptions of the process of child development, the provision of additional human resources can trigger changes in teachers' time use that may counteract any positive direct impact of these resources. However, training teachers may change their perceptions of the importance of different inputs, lead to improvements in the efficiency of how they utilize their and TA's time, and, correspondingly, deliver improvements in child development.

At the broadest level, we view this paper as furthering our understanding of how to ensure that large-scale, government-run ECE services targeted at disadvantaged groups are of sufficient quality to deliver the significant and lasting benefits that smaller programs implemented under carefully controlled conditions have been shown to have (Heckman et al. 2010; Engle et al. 2011; Heckman, Pinto, and Savelyev 2013). Our design enables us to evaluate rigorously the impact of the Colombian government's approach to quality improvement as it was, in practice, implemented nationwide. This means that these estimates bypass concerns about whether program impacts estimated through randomized controlled trials (RCTs) will hold when programs are scaled (Heckman 1992; Deaton 2010; Banerjee et al. 2017; Bold et al. 2018). Importantly, we also provide evidence on a concrete, scalable way in which the government could improve the program to deliver significantly better outcomes for children at little extra cost. This has relevance beyond Colombia, as governments in developing countries are increasingly facing the challenge of how to improve existing ECE services rather than how to start them up.

Our paper contributes to several, more specific strands of the literature. The first looks at whether and how providing schools and preschools with additional resources improves the quality of the education they deliver (see Glewwe et al. 2011 and Evans and Popova 2016 for reviews). In particular, we examine a common approach to increasing resources: providing preschools and primary schools with TAs. There is recent evidence from LMICs suggesting that the addition of TAs can generate significant benefits for primary school children when the TAs have clearly assigned tasks for which they are adequately trained (Banerjee et al. 2007; Duflo, Kiessel, and Lucas 2020). This is in contrast to older evidence from a series of evaluations of the US Tennessee Student/Teacher Achievement Ratio project. Here, while researchers found that reducing class size had significant positive impacts (especially at kindergarten level), adding TAs had no discernible impacts (Hanushek 1999; Krueger 1999; Krueger and Whitmore 2001); this may have been because these TAs were expected to perform activities they were not trained to do (Gerber et al. 2001). Indeed, Agostinelli, Avitabile, and Bobba (2023) highlight the crucial role that training of auxiliary educational professionals can

play: when mentors in Mexico had only the standard government training, their addition did not improve educational outcomes, but when they had received enhanced training focused on the precise set of tasks they were meant to perform, educational benefits followed. Our evidence suggests that TAs lacking clearly defined tasks and teachers having scope to endogenously react to the increase in TAs by reducing their own effort may have contributed to the null effect of the Colombian government's flagship program.

Second, we contribute to the literature on the impact of teacher professional development programs. Findings in the (relatively small) US literature on the impact of adding teacher professional development programs to existing ECE programs have been very mixed (Joo et al. 2020). This is also the case for the handful of rigorous studies in LMIC contexts. While there is evidence that children benefit from being in higher-quality classrooms and with higher-quality teachers in preschool (Araujo et al. 2016), two evaluations of teacher training and professional development programs in very different contexts (Chile and Malawi) found that, despite evidence of improvements in teachers' practices, there were no improvements in child development (Yoshikawa et al. 2015; Özler et al. 2018). These studies suggest that this might be due to the low intensity of the training, meaning that improvements to teachers' practices were too modest to substantially affect child development. This hypothesis is consistent with a study by Wolf (2018) of a kindergarten teacher training program in Ghana, which found that an intensive training program led to both substantial improvements in classroom practices and small improvements in child development. Our results offer further encouraging evidence on the potential of teacher training programs to change ECE teaching practices in ways that translate into improvements in children's outcomes, highlighting the importance of future research on what are the critical ingredients of effective preschool teacher training programs.

The rest of this paper is organized as follows. Section II provides details about the study setting and the interventions that we evaluate. Section III presents the study design and empirical strategy we use. In section IV, we describe our outcome measures and how we use them in the analysis. The estimates of the main impacts are presented in section V, alongside robustness checks and heterogeneity analysis. In section VI, we explore potential mechanisms using novel data on teacher time use and quality of the classroom learning environment. Section VII concludes.

II. Setting and Interventions

The programs we evaluate were aimed at improving the quality of Hogares Infantiles (HIs), which are partially subsidized government preschools for

children between the ages of 18 months and 5 years from low-socioeconomic-status families.² HIs serve children whose parent(s) are working and who are therefore at risk of inadequate childcare. The HI program is the oldest public center-based childcare provider in Colombia, and the centers have enrolled an average of 125,000 children per year over the past decade. At the time of this study, there were 1,008 HIs across the country.

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. These teachers have a significant amount of autonomy over what they do with the children in the classroom and how they utilize available resources. The teachers in our sample (described below) reported doing a wide range of activities with the children over the course of an average week, from providing them with basic care such as feeding, cleaning, and putting them down for naps to overseeing free play and implementing group and individual learning activities. The most frequent activities included attending to children's physical care needs, engaging children in conversation, and singing. These teachers have a high workload: the average teacher reported working 1.5 hours longer than their contracted hours each week.

In 2010, the government of Colombia started a comprehensive strategy to improve early childhood policies with a US\$1.28 million program called *De Cero a Siempre* ("From Zero to Forever"; see Bernal et al. 2019; Bernal and Ramírez 2019). In 2011, as part of this strategy, the improvement of the HI program was announced and the new intervention was labeled "Hogares Infantiles Mejorados" ("Improved HIs"; HIM). Specifically, HIs were given a substantial amount of additional resources, mainly for hiring new staff. The single largest pot of money was earmarked for hiring teaching assistants (TAs) to support the teachers. Prior to this program, TAs were rarely used in HIs. Government guidance suggested that, with the new money provided by HIM, HIs should aim to hire one full-time TA for every 50 children. In addition, the funds included an allocation for hiring a full-time socioemotional expert and nutritionist for every 200 children.³ While the additional funds were provided with guidance on how to use them, in practical terms HIs had complete autonomy over this, since there were no monitoring mechanisms in place. In spite of this autonomy, we show in the next section that compliance with the guidance was high.

We worked with the government to embed an RCT into the initial HIM rollout. To this end, a random subset of HIs were wait-listed to receive the

² Occasionally, HIs take children as young as 6 months when it is judged 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.

³ This paper focuses on impacts on child development. In table B.14 (tables A.1–D.1 are available online), however, we document that we see no evidence that either program had impacts on nutritional outcomes once we have corrected for multiple hypothesis testing.

program 18 months later. Additionally, there was interest from Fundación Éxito and the Colombian National University in offering a teacher professional development training program alongside the resources provided to HIs by the government HIM program. We therefore added an arm to the RCT in which HIs received the resources through HIM and teacher professional development training through Fundación Éxito. The training program was developed by Fundación Éxito in partnership with the Colombian National University. The curriculum covered modules on the process of child development between the ages of 18 and 36 months; the importance of different inputs for child development, including, for example, the use of art, music, and body language; and pedagogical strategies for providing these inputs. In response to a concern that teachers allocated too much class time to basic caregiving activities, the program placed strong emphasis on the importance of focusing on activities that promote child development during class time and best practice in implementing these.

The FE training program was delivered over the course of 13 months through three components: (i) 16 3-hour sessions spread over the 13 months in which the group of teachers were physically together and the instructor connected via videoconferencing software; (ii) 3 hours per week of video tutoring sessions in which participants worked with their tutors online on developing and refining classroom activities; and (iii) on-site coaching where instructors carried out one classroom observation of participating teachers to provide specific feedback on their content and pedagogical methodology. It is important to note that implementation of training via videoconferencing is a key feature for the scalability of this program in contexts where appropriate technology is available through greatly reducing costs and logistical complexity. The program was offered for free, but participating teachers incurred costs of transportation to monthly sessions (which often could not take place in the HIs themselves because of the lack of a reliable internet connection), required internet access for the tutoring sessions, and needed materials for preparation of new activities. In addition to this training, teachers as well as parents were offered reading workshops in which they were trained on how to read with children, and training centers received books and book bags to distribute among participants.⁴

The HIM program cost the government a substantial amount, a 30% increase in per-child expenditure relative to the “business-as-usual” model without enhancement, which amounted to extra expenditure of \$300 per

⁴ We find no impact on any indicator of reading routines in the home. See table B.13 for details. The FE program also included a nutritional improvement component that aimed to increase calorie provision by 15% above the 60% of daily requirements already provided by HIs. In table B.14, however, we document that we see no evidence that either program had impacts on nutritional outcomes once we have corrected for multiple hypothesis testing.

child per year. Precise cost calculations of the FE component are more challenging. However, imputations based on reasonable assumptions suggest that its cost is a small fraction of the cost of the HIM program: following an up-front investment of around \$34 per child (\$5,827 per HI) for initial training, we estimate the cost of refreshers and training for new starters to be about \$13 per child per year (\$2,206 per HI). See appendix A (apps. A–E are available online) for details of calculations.

III. Study Design and Empirical Strategy

We designed a three-armed cluster RCT around the national rollout of the HIM program in order to assess the effects of HIM alone and of the augmented version, HIM+FE. The study took place in the eight largest cities in Colombia: Bogotá, Cali, Medellín, Barranquilla, Bello, Palmira, Itagüí, and Soledad. These are also the cities with the largest number of HIs. Forty HIs were randomized into each of the three arms: (i) HIM, in which preschools received the government quality improvement program; (ii) HIM+FE, in which preschools received the teacher professional development training enhancement in addition to the HIM program; and (iii) a pure control group for which the implementation of HIM was delayed. This design allows us to test whether the government improvement program had an impact on children attending the upgraded centers relative to those in the “business-as-usual” HIs, evaluate the full impact of the HIM+FE program relative to “business-as-usual” HIs, and test whether adding the FE component represents an improvement over and above the government upgrade.

To select the 120 study HIs, we first obtained global positioning system coordinates for all of the HIs in the eight study cities (248 in total). 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. To be eligible, HIs had to have at least 15 children in our target age range (18–36 months at baseline). Within each triplet of eligible HIs, we randomly assigned one to the pure control group, one to the HIM treatment group, and one to the HIM+FE treatment group. Randomization and sample selection were carried out over November–December 2012.

On average, the HIs in the sample had 48 children between the ages of 18 and 36 months; we drew a baseline sample of 15–17 children per HI from this group.⁵ Baseline data were collected between March and May 2013. While HIs assigned to HIM and HIM+FE had already begun to

⁵ We included all of the children in HIs where there were 15, 16 or 17 children in the target age range. If there were more than 17 children in the target age range, we randomly selected 17.

make preparations for the HIM upgrades at the point of baseline, we do not see any imbalances that might be evidence of the program already having effects on child development. The total baseline sample consisted of 1,987 children (663 in HIM centers, 663 in HIM+FE centers, and 661 in control group HIs). End line was conducted 18 months later, in October and November 2014. Our aim was 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. Some of the child development assessments (our key outcome measures) were unsuitable for children below the age of 48 months. Therefore, our main analysis sample comprises only children above 48 months at end line who were thus eligible for all assessments. We return to this issue in section V.B

A. *Balance and Attrition*

Attrition was relatively low. We completed some end-line child development assessments for all but 155 (7.8%) of the 1,987 children in the baseline sample. As discussed above, we exclude the 753 children who were under 48 months at the time of the assessments from our main analysis sample, since these children were not eligible for the complete set of child development assessments. This leaves us with 1,074 children with complete assessment data in our main analysis sample. The attrition rate among children who were 48 months or over at the time that assessments were held at their HI was 6.8%. In both the “extended sample,” which includes younger children for whom we have incomplete assessment data, and the main analysis sample, attrition was not related to treatment assignment (table B.1).⁶

Table 1 shows baseline characteristics of our main analysis sample, split by treatment assignment. On all sociodemographic characteristics other than gender, the sample appears well balanced. While the control group is slightly more female than either treatment arm, we do not see this imbalance reflected in baseline child development, and we control for gender in all analysis. The sample is well balanced in children’s problem-solving, language, communication, and socioemotional skills. We do see slight imbalances in fine and gross motor skills in which the HIM group appear to have slightly higher skills at baseline. We control for all domains of baseline child development (including fine and gross motor skills) in our main estimates of treatment effects on child development.⁷

⁶ In our robustness analysis in sec. V.B we show we obtain similar results when examining either sample.

⁷ When examining the coefficients on these control variables (in table B.6), it is reassuring (given these slight imbalances) to note that baseline motor skills seem unimportant in predicting end-line cognitive and socioemotional development. In contrast, baseline measures

The majority of children (72.2%) continued attending the same HI throughout the study period; by end line, 9.2% were enrolled in a different HI (mostly an HI not in the study sample), 13.1% were enrolled in a different public or private childcare service, and 5.5% were not enrolled in any type of childcare service. The probability that children remained in the same HI was not affected by treatment status. We aimed to survey all children in the baseline sample, irrespective of whether they were in the HI they attended at baseline.

B. Compliance

The HIM program instructed HIs to hire one full-time TA for every 50 children, as well as a full-time socioemotional expert and a nutritionist for every 200 children. While we do not directly observe either the amount of money provided to HIs through the HIM program for the extra hiring, or how that money was spent, we can deduce both from our data. We use data on number of children in a given HI to first impute the total extra budget allocated to each HI through the HIM program to spend on hiring new staff. We then use personnel data, including data on salaries for TAs, nutritionists, and socioemotional experts, collected at baseline and end line to compare the actual spending on additional staff salaries with what was expected. This exercise suggests that, on average, compliance was high, with more than 70% of the money allocated for hiring being spent in this way across the two treatment arms.

At end line, preschools in the HIM and HIM+FE arms both had an average of 0.94 TAs employed for every 50 children (table B.2) with almost all TAs working full-time. This result falls just short of the HIM target of 1 TA per 50 children. On average, in preschools allocated to HIM and HIM+FE there were, respectively, 0.47 and 0.45 TAs for every teacher. Almost all preschools in these treatment arms had also hired a nutritionist and socioemotional expert (indeed, 90% had hired at least one of each type of professional) although many of these staff were working part-time (table B.2). Salary data suggest that HIM hiring targets for these professionals might have been overly optimistic, given actual market wages, leading to many nutritionists and socioemotional experts being employed only part-time.

The FE teacher professional development training took place between June 2013 and June 2014. HI directors nominated two or three teachers per treated HI to participate, with some additional teachers from the same HIs selected to replace teachers who were not able to attend all

of language, problem-solving and communication skills are highly predictive of end-line outcomes.

TABLE 1
BASELINE SOCIODEMOGRAPHIC CHARACTERISTICS AND CHILD DEVELOPMENT
BY RANDOMIZATION STATUS FOR ANALYSIS SAMPLE

	Control	HIM	HIM+FE	HIM vs. control <i>p</i> -Value	HIM+FE vs. control <i>p</i> -Value	HIM vs. HIM+FE <i>p</i> -Value	<i>N</i>
Male	.456 (.499)	.552 (.498)	.522 (.500)	.007	.079	.468	1,074
Age (months)	32.98 (2.120)	32.77 (2.179)	32.73 (2.200)	.287	.166	.836	1,074
Household income (million COP)	1.33 (.774)	1.34 (.778)	1.34 (.796)	.901	.959	.957	1,074
Mother's education (years)	12.63 (2.776)	12.37 (2.601)	12.66 (2.579)	.301	.864	.131	1,064
Father's education (years)	12.01 (3.041)	11.98 (3.116)	12.13 (3.073)	.911	.699	.541	1,003
Household size	3.385 (1.697)	3.477 (1.629)	3.217 (1.542)	.517	.170	.061	1,074
ASQ-3 communication	63.95 (19.77)	65.86 (20.84)	64.33 (20.13)	.326	.843	.424	1,074
ASQ-3 gross motor	62.22 (21.67)	66.22 (20.89)	64.53 (20.03)	.059	.178	.417	1,074
ASQ-3 problem solving	57.63 (19.51)	59.40 (20.35)	58.71 (19.20)	.414	.589	.721	1,074
ASQ-3 personal social	57.86 (18.59)	60.49 (18.59)	59.01 (18.37)	.151	.508	.330	1,074
ASQ-3 fine motor	46.98 (20.09)	51.50 (20.81)	46.56 (19.73)	.070	.875	.037	1,074
MacArthur- Bates language	66.16 (24.09)	67.68 (24.03)	66.49 (23.41)	.566	.911	.647	1,074
ASQ:SE	56.09 (21.42)	53.29 (19.73)	54.61 (20.65)	.137	.490	.466	1,074
Observations	353	384	337				

NOTE.—Baseline means (SD) by treatment status for children included in the analysis sample (i.e., all children with complete child development assessment data at end line). Two-sided *p*-values are estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). ASQ-3 child development scores are the raw scores from the five subscales of the ASQ-3: communication, gross motor, problem solving, personal social and fine motor. Socioemotional score is the raw scores from the ASQ:SE. MacArthur-Bates language is the raw score from the MacArthur-Bates CDI. Child development measures are described in sec. IV.B. COP = Colombian pesos.

of the sessions or who dropped out. Administrative records indicate that 114 (out of 309) teachers in the 40 HIs assigned to HIM+FE started the training. Of these, 99 teachers (or 87%) were certified as having completed it. Although the training was designed for teachers, in rare cases other staff also participated, including TAs, directors, and other senior staff. We do not have information on numbers or characteristics of teachers who were nominated by the center director or which nominated teachers enrolled. We are also not able to link the teachers and TAs in our sample to FE records of those who enrolled. We therefore are not able to identify children in the HIM+FE sample who were taught by a teacher who received FE training.

C. Empirical Strategy

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

$$Y_{ilm} = \beta_0 + \beta_1 T_{lm}^{\text{HIM}} + \beta_2 T_{lm}^{\text{HIM+FE}} + X_{ilm}\gamma + \epsilon_{ilm} \quad (1)$$

where Y_{ilm} is the outcome of interest for child i , in preschool l , in triplet m ; X_{ilm} is a prespecified set of control variables added to improve efficiency; and ϵ_{ilm} is the random error term. We allow for correlation between errors from observations belonging to the same sampling triplet (de Chaisemartin and Ramirez-Cuellar 2020).

Prespecified baseline controls for child-level outcomes include the child's age, age squared, gender, a set of city dummies, and child development measured at baseline. We discuss how outcomes were measured at baseline and end line in section IV.B. For teacher- and classroom-level outcomes we control for the baseline level of the relevant variables averaged at the HI level.⁸ For classroom-level outcomes we also control for the average age of the kids in the class.

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 and two-sided p -values for testing the null hypothesis that the treatment effect in question is zero using a cluster bootstrap with 1,000 iterations.

⁸ Using averages allows us to control for these variables even for teachers who began working at the center since baseline and thus were not in our original baseline sample. In table B.12, we show that our results are robust to including only teachers who were present in the HI at baseline.

We cluster at the triplet level (de Chaisemartin and Ramirez-Cuellar 2020).

When we test the same hypothesis (i.e., the difference between any two treatment arms) on multiple conceptually similar measures of child development, we also present q -values that are adjusted for multiple testing across these outcomes. To do this, we use the stepwise procedure described in List, Shaikh, and Xu (2019, building on Romano and Wolf 2005, 2010), which provides balanced asymptotic control of the family-wise error rate. In running the procedure, we use the cluster bootstrap described above, studentizing by the bootstrapped 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 power of the tests compared with classical methods.

IV. Outcomes and Measurement

Measuring the variables we are interested in—that is, different dimensions of child development and the features of the preschool environment that are important for child development—is not trivial. We collected rich measures of child development, the classroom environment, and teaching practices. In this section, we describe these measures and how we use them.

In section V, we start by presenting estimates of impacts on measures scored using the standard algorithms recommended by the test publishers. Additionally, we follow the literature in using structural measurement models to summarize the information contained in our measures efficiently. Here, we first outline, in section IV.A, the measurement models we use and how we estimate the latent factors of interest in the analysis. In section IV.B, we then discuss the specific measures of child development in our analysis and how we use them to construct estimates of latent factors for (1) child cognitive development and (2) child socioemotional development. In section IV.C, we do the same for measures relating to the mechanisms through which the two interventions may have shifted child outcomes: (1) teachers' overtime hours, (2) teachers' participation in "learning activities" within the classroom, (3) teachers' participation in "personal care" activities, (4) TAs' participation in learning activities, (5) TAs' participation in care activities, and (6) the quality of the classroom learning environment as directly observed by a psychologist.

A. Measurement Model

We adopt the increasingly common approach of using a structural measurement model to construct estimates of underlying latent factors capturing

each of our outcomes (see, e.g., Cunha, Heckman, and Schennach 2010; Heckman, Pinto, and Savelyev 2013; Attanasio et al. 2020; Heckman, Liu, Lu, and Zhou 2020; Heckman and Zhou 2022; Agostinelli, Avitabile, and Bobba 2023). These techniques combine the information contained in the available data efficiently. Furthermore, they model measurement error directly. This allows for the estimation of treatment effects scaled relative to the true variance of the underlying construct in the control group, uncontaminated by variability induced by measurement error. These benefits of estimating a within-sample structural measurement model lead to improvements over adopting official scoring algorithms, especially if the official algorithms were developed using data from a population very different from the study population.⁹

Since we have rich item-level data capturing the binary or ordinal responses of children, parents, and teachers to each item within each instrument, we opt for a measurement model based on item response theory (IRT). These methods have a long history in psychometrics (Van Der Linden and Hambleton 1997) and are increasingly being used by economists (e.g., Das and Zajonc 2010; Heckman et al. 2020; Singh 2020; Heckman and Zhou 2022). While linear factor models model multiple continuous test scores as depending linearly on underlying unobserved factors (e.g. Cunha, Heckman, and Schennach 2010; Heckman, Pinto, and Savelyev 2013; Attanasio et al. 2020; Agostinelli, Avitabile, and Bobba 2023), IRT models use nonlinear linking functions (such as logit and ordered-logit models) to map responses to discrete items onto unobserved latent factors. Past work has shown that estimating underlying factors directly from individual binary or ordinal item responses yields performance gains if items vary substantially in their difficulty and discrimination power (Van Der Linden and Hambleton 1997; Heckman et al. 2020). We expect this to be the case in our setting since our assessment items are designed to get more difficult as the test progresses.

Specifically, let θ_{id} represent i 's factor of interest in domain d , where d can be cognitive development, socioemotional problems, learning activities, care activities, or directly observed classroom quality, and where i represents the individual child, teacher, TA, or classroom. However, θ_{id} is not observed directly. Instead, observable item responses, y_{ijds} are noisy measures of the latent factor θ_{id} . Our main results make a number of key assumptions about these latent factors and the mapping from latent factors to item responses:

⁹ For example, the official scoring algorithm provided with the Woodcock-Muñoz tests, which we use to measure cognitive development, converts patterns of responses into standardized scores using parameters estimated using a measurement model on a norming sample comprised of 1,413 Spanish-speaking children from the United States, six Latin American countries, and Spain (Schrang et al. 2005). This norming sample is likely to differ substantially from our sample.

- A1. There is a dedicated and unidimensional measurement system for each domain.—In other words, we assume that responses for items in domain d may depend on the underlying unidimensional factor θ_{id} but not on $\theta_{id'}$ for $d' \neq d$.
- A2. The noise in the mapping of latent factors to item responses is independent across respondents and across items within a domain.
- A3. Underlying latent factors are normally distributed in the control group.—As location and scale normalizations, we impose a zero mean and unit variance in the control group.¹⁰ Our approach does not assume normality in the treatment groups.
- A4. Treatment does not affect the mapping from latent factors to item responses.

In section V.B, we show evidence that supports assumption A1: a single unidimensional factor for each domain in our application and sample. However, we note that Heckman et al. (2020) extend these same methods to allow for multidimensional factors, potentially correlated, entering each measurement system. In section V.B, we also show that our results are not sensitive to relaxing assumptions A2 and A3. Furthermore, Heckman, Pinto, and Savelyev (2013) and Heckman et al. (2020) point out that the mapping between children's ability and certain item responses could be affected by exposure to an intervention. We thus explore the validity of assumption A4 by testing for measurement invariance item by item in appendix D and discuss the findings in section V.B. Overall, we find no evidence that either treatment altered this mapping. We thus impose invariance for our main analysis but show two additional robustness checks in appendix D.

Over and above these main assumptions, we make specific functional form assumptions about the mapping from latent factors to item responses. Depending on the nature of each item, we use one of three different specifications. First, we have binary items where it is conceptually possible for the correct response to be "guessed." For example, a child with a low level of development may still guess the correct answer to a difficult question. We model these items using a three-parameter "guessing" specification (Birnbaum 1968) to describe the probability that i correctly answers item j :

$$\Pr(y_{ijd} = 1 | \theta_{id}) = g_{jd} + (1 - g_{jd}) \frac{\exp(\alpha_{jd} + \beta_{jd}\theta_{id})}{1 + \exp(\alpha_{jd} + \beta_{jd}\theta_{id})}. \quad (2)$$

¹⁰ As we are not interested in explicitly estimating the process of child development over time (unlike, say, Agostinelli and Wiswall 2016) but only seek to use baseline values as control variables, we normalize the relevant factor at each wave (baseline and end line).

In this set-up, α_{jd} represents an item j 's difficulty—the higher the value of α_{jd} , the easier an item is; β_{jd} represents its discriminatory power and governs the rate at which the probability that the item is answered correctly changes with the underlying factor; and g_{jd} is the “pseudoguessing parameter” and is the asymptotic probability of i choosing correctly as $\theta_{id} \rightarrow -\infty$.¹¹

Second, we have some binary items where it is not conceptually possible to guess the correct answer, such as a psychologist's report of whether or not they observed certain indicators of classroom quality. For these, we use a standard two-parameter IRT model that is the same as above but restricts the guessing parameter (g_{jd}) to 0.

Third, we have some items that have three or more ordinal response categories. For instance, one of the child development assessments records how many words in a particular category a child can name, and our measures of teachers routines are based on the number of days on which a teacher carried out a particular activity during the last week. For these, we use a “graded” model that models the probability of i having a response of more than k as an ordered logit:

$$\Pr(y_{ijd} \geq k | \theta_{id}) = \frac{\exp(\alpha_{jkd} + \beta_{jd}\theta_{id})}{1 + \exp(\alpha_{jkd} + \beta_{jd}\theta_{id})}. \quad (3)$$

We drop items with little variation in the estimation. Specifically, we drop binary items where more than 90% of responses take a given value.¹² We estimate the measurement models by maximum likelihood, using an expectation-maximization algorithm and using Gauss-Hermite quadrature to approximate the integral over the unobserved latent factor. We follow the literature in adopting unbiased estimators for each i 's underlying factor. While for linear models, Bartlett scores (Bartlett 1937) provide unbiased estimates, for our nonlinear setup we obtain unbiased estimates for

¹¹ In practice, estimating our most general model leads to us estimating that some guessing parameters (the g_j 's) are very close to the lower bound of zero. Intuitively, this just means that we estimate that a very low-ability child has a negligible chance of guessing the correct answer. However, this can lead to problems in inverting the estimated information matrix. Therefore, we follow a multistep procedure where we first estimate the unrestricted model. We find all items with an estimated guessing parameter greater than 0.05. We then reestimate the model allowing all of these items to have a freely estimated guessing parameter but restricting the guessing parameter on all other items to be zero. Table C.1 shows which items are estimated to have positive guessing parameters. In table B.7, we show that our results are not sensitive to this choice. Allowing all items to have positive guessing parameters produces very similar results, but the estimated information matrix is not invertible.

¹² We do this in order to ensure that the associated information matrix is invertible. In table B.7, we show that our results are not sensitive to this choice: our main estimates remain almost identical if we change this threshold to 99%.

each θ_{id} by maximizing the likelihood of observing the realized response patterns conditional on the estimated parameters. When we estimate treatment effects on these predicted scores, we bootstrap the entire procedure (including reestimating the measurement system on every bootstrapped sample) to account for noise arising from the measurement system.

When estimating the measurement system, we assume that the underlying factors are normally distributed. However, this assumption might be violated in the presence of treatment effects and heterogeneity in these. Because of this, we estimate the measurement system on the control group only and use the estimated parameters of the measurement system to derive estimates of the relevant latent factors in the treatment group.

B. Child Development

We now turn to the specific measures of child development in our analysis and how we use them to construct estimates of latent factors for child cognitive and socioemotional development. Child development is a multidimensional construct, as discussed, for instance, in Cunha, Heckman, and Schennach (2010), Attanasio, Meghir, and Nix (2020), Attanasio et al. (2020), and Heckman et al. (2020). Furthermore, preschool has been shown to affect various dimensions of children's development (Berlinski, Galiani, and Gertler 2009; Datta Gupta and Simonsen 2010; Chetty et al. 2011; Heckman, Pinto, and Savelyev 2013; Araujo et al. 2016; Kline and Walters 2016). Therefore, at both baseline and end line, we used a range of child development assessments that sought to capture children's development across different domains. The measures we used at end line were richer than those used at baseline. This was due to a combination of cost considerations and the fact that the emphasis of the study is on estimating treatment effects on end-line child development, so that baseline measures are primarily useful for checking for balance and increasing the precision of estimated effect sizes.

1. Baseline

At baseline, we administered all five subscales of an extended version of the ASQ-3 (Ages & Stages Questionnaires, third edition) to measure communication, gross motor, problem-solving, personal social, and fine motor skills (Squires, Bricker, and Twombly 2009);¹³ the MacArthur-Bates Communicative Development Inventories (CDI; Jackson-Maldonado et al.

¹³ The standard ASQ-3 comprises age-specific questionnaires containing six questions for each subscale. We extended each age-specific questionnaire by adding the last three

2003; Jackson-Maldonado, Marchman, and Fernald 2013) to measure language development; and the ASQ:SE (Ages & Stages Questionnaires: Social-Emotional; Squires, Bricker, and Twombly 2002) to measure socio-emotional development. These are all parental report instruments; that is, parents are asked to report on the development of their children.

For each of these eight baseline assessments, we have a series of binary items indicating the parents' assessment of whether their child can do a specific task.¹⁴ For each assessment separately, we combine items using two-parameter IRT model described in section IV.A. Table C.3 presents the parameter estimates for these measurement models alongside estimated confidence intervals. Our estimates show that the vast majority of items have discrimination parameters that are significantly greater than zero; in other words, they are informative of the underlying factors.

In order to control for baseline child development in the most flexible manner, we include the full set of factor scores estimated using the seven baseline assessments. In robustness analysis (table B.8), we show that controlling for baseline child development using raw scores, rather than IRT scores, makes no difference to our estimates.

2. End Line

We have end-line data for seven child development assessments, each designed to capture a different dimension of child development: (1) fluid reasoning, (2) memory for words, (3) expressive language, (4) receptive language, (5) school readiness, (6) inhibitory control, and (7) socio-emotional development.¹⁵ Assessments 1–3 comprise the relevant scales

nonoverlapping items in each subscale from the age-specific questionnaire below and the first three nonoverlapping items in each subscale from the age-specific questionnaire above. This was to ensure that the instrument had sufficient information over the entire support of baseline child development. Because questionnaires differ depending on the age of the child, not every indicator is answered for every child in the ASQ-3. However, there is strong overlap by age, which allows us to use our IRT model to estimate a single factor for each subscale. For these items, there were three possible answers respondents could give: "never," "sometimes," and "always." However, we found that parents very rarely chose "sometimes." We therefore convert these to binary items by splitting above and below the mean value (which is equivalent to combining the "sometimes" responses with the category with the next-fewest responses).

¹⁴ The MacArthur-Bates CDI has separate lists of words for children above and below 30 months of age. We score both in separate IRT models. When controlling for baseline child development, we control for both factors simultaneously, replacing undefined values by the average score for that assessment and adding a dummy indicator for the assessment used.

¹⁵ We also collected measures of sound awareness and concept formation. However, these two tests were too hard for most children, many of whom did not progress past the initial few items, leaving very little information. Specifically, only 25.9% of children progressed past the first five items (out of a total of 29) in the test of concept formation (Woodcock-Muñoz-III [WM] cognition 5) and only 5.1% of children progressed past the first nine

from the WM tests of cognition and achievement (Schrank et al. 2005), which are Spanish versions of the well-known Woodcock-Johnson tests (Woodcock 1977). Receptive language was measured using the Spanish version of the Peabody Picture Vocabulary Test (Test Visual de Imágenes Peabody [TVIP]; Dunn et al. 1986), and school readiness using a shortened version of the Daberon-II (Danzer et al. 1991) that included only 70 items, which were chosen through piloting. Inhibitory control, a dimension of executive functioning, was measured using the nonverbal Pencil Tapping Task (PTT; Diamond and Taylor 1996). Finally, socioemotional development was assessed using the ASQ:SE (Squires, Bricker, and Twombly 2002). Table B.3 provides details of all assessments.

The first six measures of child development, which, broadly speaking, capture skills related to cognitive development, school readiness and language, were collected through direct assessments of children by trained psychologists, undertaken in the HIs. Given the challenges of assessing socioemotional development in young children directly, we relied on parental reports, introducing the ASQ:SE module as part of the questionnaire to the child's primary caregiver. We chose assessment tools that had previously been validated for use in Latin American populations. Most of the measures we selected had previously been used in Colombia, for instance in Bernal and Fernández (2013) and Andrew et al. (2018).

As already noted, we score these measures in two ways: in accordance with the official algorithms recommended by the test publishers and using a measurement model based on IRT. We use the scores that are not prestandardized for age in order to allow for a more flexible age gradient. To construct publisher-recommended scores, we use the *W*-scores, which are created for the WM tests using the publisher's algorithm based on IRT. For the TVIP, we use the recommended scoring algorithm to create the "raw score." The Daberon and PTT are more straightforward since all children answered all items. Hence, here we simply use the total number of correct responses. For the ASQ:SE, which is reverse scored (so higher scores mean lower socioemotional development), we follow the publisher's guidelines, assigning a score of 5 when the carer answered "sometimes" and 10 when they answered "rarely or never."

We check that our measures pass basic tests of internal validity. We find that our measures of child development at end line are strongly correlated with age, baseline child development, household wealth, and maternal education in the expected direction (see table B.4) and are strongly positively correlated with one another (see table B.5). Maternal report measures of socioemotional development show lower correlations with age, baseline socioemotional development, household wealth, and maternal

(out of a total of 18) items on the test of sound awareness (WM achievement 21). Because of this poor performance, we drop these assessments from all analyses.

education (table B.4) than the direct assessment measures. These lower correlations could be a feature of socioemotional skills or a sign that the maternal report measures are noisier measures of development.

We summarize items from all assessments measuring constructs related to cognition, language, and school readiness (assessments 1–6) into a single estimated factor using the procedure outlined in section IV.A. We label our resulting estimated factor “cognitive development.” We then summarize all items from the ASQ:SE using a separate measurement system and estimate a “socioemotional problems” factor for each child. As we discussed in section III.C, we reestimate the measurement system in every bootstrapped sample when estimating treatment effects so that our inference accounts for the fact that our outcome measures are themselves estimated.

Tables C.1 and C.2 present our parameter estimates for these measurement systems alongside confidence intervals. Importantly, for both cognitive and socioemotional development, almost all of the items appear to be informative of the underlying factor. For cognitive development, for instance, all 118 of our estimated discrimination parameters (the β_{jd} 's) are positive and only 6 out of 118 have 95% confidence intervals that contain zero. When taken as a whole, a useful summary measure of the precision of our predicted latent factors is that the standard deviation of the mean (median) across all bootstrapped samples of a given child's predicted factor score is 0.15 SD (0.13 SD) for cognitive development, where SD refers to the metric used for that factor. The corresponding figures for socioemotional problems are 0.24 and 0.19 SD, indicating that these estimates are slightly less precise.

C. Classroom Activities and Preschool Quality

We collected detailed measures of classroom activities in order to assess whether and how the interventions changed the routines and quality of instruction among HI teachers and TAs. First, we collected teacher overtime hours measured as the number of hours teachers reported working over and above their contracted hours on a typical week. Second, we collected detailed self-reported data on the type of activities teachers and TAs had performed in the classroom over the week prior to the interview (from a list of 36) and on how many days they had performed them. These questions come from the Teacher Survey of Early Education Quality (Hallam et al. 2011).

We split the teacher- and TA-reported activities into two groups. The first group comprises *care activities* which relate to basic care of children such as changing diapers, brushing teeth, and washing hands, as well as napping and feeding routines. The second group comprises *learning activities*,

such as reading stories, teaching skills, storytelling, and singing. This split is motivated by three factors. First, there is a large literature suggesting that psychosocial stimulation is a key determinant of children's development (Heckman and Mosso 2014; Attanasio et al. 2020), so we seek to separate out activities focused on delivering such stimulation. Second, FE training emphasized the importance of highly stimulating activities for children's development. And third, given that teachers are trained to deliver learning activities as part of their ECE qualification (but the TAs are not), a natural allocation of roles would be for the teachers to focus on learning activities and the TAs to focus on care activities.

We construct summary measures of each of the two broad categories of activities, separately for teachers and TAs, using the procedure described in section IV.A. Specifically, we take the number of days that teachers, or TAs, reported doing each of the care activities and adopt the graded-response specification described in equation (3). We repeat this procedure for the learning and development activities. Tables C.4 and C.5 show the full set of activities used and the estimated parameters in the measurement systems for teachers' learning and care activities respectively. Tables C.6 and C.7 show the same for the TAs' activities. These estimates suggest the measures performed well; almost all items are significantly informative about the relevant underlying factor.

In addition to these self-reported measures of teachers' and TAs' activities, we measured the quality of teaching activities through direct observation of the teachers 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 and has been used extensively across a wide range of cultural and economic contexts. It has been shown to be predictive of child gains in cognitive (Burchinal et al. 2000; Peisner-Feinberg et al. 2001) and socioemotional development (Sylva et al. 2006). The ECERS-R was carried out by psychologists, who were trained for three weeks; each classroom observation lasted at least half of a school day. Observations were carried out only when the teacher was present in the classroom and teaching. Because of logistical and budgetary constraints, we conducted ECERS-R in only 172 of the 847 classrooms in our sample.¹⁶

¹⁶ The subsample was chosen as follows. At baseline, we randomly chose 216 classrooms attended by study children in 54 HIs selected randomly, stratifying by city, in which to measure classroom quality using either the ECERS-R (suitable for classrooms with children over 2 years of age, 60% of classrooms) or the Infant/Toddler Environment Rating Scale, Revised (ITERS-R; the 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 (because study children had moved on from their baseline classrooms). This resulted in observations in 172 classrooms with children older than 2 years where we carried out the ECERS-R and 39 classrooms with children aged 0–2 where we

The ECERS-R is comprised of 43 individual items, each measuring a different aspect of quality—for example, “encouraging children to communicate.” We exclude items related to the “space and furnishings” subscale since our interventions did not target the physical quality of the classroom environment. Instead, we take all items contained in the other six subscales—“personal care routines,” “language-reasoning,” “activities,” “interactions,” “program structure,” and “parents and staff”—that relate to the quality of teaching processes within the classroom. Each item comprises several indicators. We take all of the indicators that were due to be answered in all observations and again summarize them using the measurement model described in equation (2).¹⁷ Table C.10 presents parameter estimates from this measurement model and suggests that almost all items load significantly onto the underlying factor.

V. The Impacts of HIM and HIM+FE on Children’s Development

This section presents estimates of the impacts that HIM and HIM+FE had on child development. We present estimates of the average impacts, as well as evidence on how the impacts differ by observed characteristics of children and their families.

Table 2 reports estimates of the impacts of the HIM and HIM+FE programs on child development measures scored according to the publishers’ recommended algorithms. Table 3 then reports impacts on estimated factor scores, which combine all items from our measures of cognitive and socioemotional development into summary factors, as described in section IV. Estimation using these factor scores has the advantage of using information contained in each assessment more efficiently. It also generates impact estimates that are scaled by the true variance of the underlying factor in the control group, uncontaminated by measurement error induced variability.

carried out the ITERS-R. We dropped the 39 ITERS-R classrooms from our classroom analysis because the sample is too small to be analyzed independently and they cannot be linked to ECERS-R classrooms because of a lack of common items.

¹⁷ Each item is formed of around 10 subitems 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 because of stopping rules that resulted in a high number of nonrandom missing values for items in the “minimal,” “good,” and “excellent” categories. We therefore use only items from the “inadequate” category in our analysis. While this overcomes the challenge posed by missing data, it implies that the subitems that make up our quality measures are informative on the absence of poor practices rather than the presence of good ones.

To increase the sample size for estimating the measurement system parameters, we pool ECERS-R measures from baseline and end line, giving a total sample of 296 observations.

TABLE 2
IMPACTS ON CHILD DEVELOPMENT ASSESSMENTS

	Fluid Reasoning (1)	Memory for Words (2)	Expressive Language (3)	School Readiness (4)	Receptive Language (5)	Inhibitory Control (6)	Socioemotional Problems (7)
HIM	-.083 (.462)	1.474 (2.992)	-.749 (1.366)	.283 (.833)	-.447 (1.510)	.253 (.367)	.723 (2.687)
<i>p</i> -Value	.877	.632	.575	.724	.330	.502	.800
<i>q</i> -Value	.971	.961	.961	.971	.854	.961	.971
HIM+FE	.918* (.487)	4.621 (2.952)	2.407* (1.261)	1.883*** (.671)	.737 (1.241)	.360 (.412)	-1.920 (2.754)
<i>p</i> -Value	.057	.115	.057	.003	.550	.384	.476
<i>q</i> -Value	.261	.372	.261	.028	.759	.759	.759
Difference	1.000** (.410)	3.147 (2.448)	3.156*** (1.104)	1.601** (.717)	2.184* (1.216)	.107 (.336)	-2.643 (2.345)
<i>p</i> -Value	.014	.185	.004	.030	.076	.771	.257
<i>q</i> -Value	.073	.474	.026	.113	.249	.771	.474
Observations	1,073	1,073	1,073	1,074	1,074	1,074	1,074
Control mean	486.31	464.31	460.13	49.84	33.54	7.14	58.30
Control SD	5.36	28.71	16.43	10.14	15.15	4.45	24.80

NOTE.—Two-sided *p*-values are estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations); *q*-values 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 (2019). Clustered standard errors (bootstrapped) are in parentheses. All estimates control for age, gender, city effects, and baseline scores for MacArthur-Bates CDI and each subscale of the ASQ-3 and ASQ:SE. All measures are scored using algorithms recommended by their publishers, as described in sec. IV.B.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

TABLE 3
IMPACTS ON COGNITIVE AND SOCIOEMOTIONAL FACTOR SCORES

	Cognitive development (1)	Socioemotional problems (2)
HIM	-.008 (.087)	-.014 (.155)
<i>p</i> -Value	.925	.933
HIM+FE	.161** (.073)	-.155 (.174)
<i>p</i> -Value	.025	.382
Difference	.169*** (.060)	-.141 (.148)
<i>p</i> -Value	.005	.341
Observations	1,074	1,074

NOTE.—Two-sided *p*-values are estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) are in parentheses. We reestimate the measurement system on each bootstrapped sample. All estimates control for age, gender, city effects, and baseline scores for MacArthur-Bates CDI and each subscale of the ASQ-3 and ASQ:SE. All factors are scaled so that the underlying latent factor has a mean of 0 and SD of 1 in the control group. All factors are constructed as described in sec. IV.B.

** *p* < .05.
*** *p* < .01.

The first row in tables 2 and 3 shows estimates of the intent-to-treat impact of HIM improvements relative to children in preschools with no improvements (pure control). The second row shows impacts of HIM+FE relative to children in preschools in the pure control group. The final row shows the impact of adding the FE component to the HIM program (i.e., the difference between the HIM and HIM+FE programs). Table B.6 shows estimated coefficients on the control variables.

A. *Child Development*

Columns 1–6 of table 2 show estimates of the impacts of the interventions on children’s performance in each of the cognitive-development assessments, scored using the algorithms recommended by the publishers. Column 1 of table 3 shows the results for the single factor representing cognitive development derived from all items from these six measures.

We see no evidence that the HIM program led to an improvement in children’s performance on any of the cognitive assessments. The lack of a significant impact of the HIM intervention on children’s cognitive development is confirmed by results in column 1 of table 3, where we find no impact of HIM on the cognitive-development factor.

However, we find evidence that HIM+FE did improve children’s performance in several of the cognitive assessments. We see evidence of a

treatment effect from the combined intervention for three assessments, in particular those measuring fluid reasoning, expressive language, and school readiness. When examining the additional effect of FE over and above HIM, reported in the “difference” row, we see statistically significant improvements across four measures: the three measures listed above as well as the measure of receptive language. These patterns are reflected in an overall positive impact of the HIM+FE program on the child cognitive-development factor, shown in table 3. We estimate that the HIM+FE program led to improvements of 0.16 of a SD relative to the pure control group, with a p -value of .025. The “difference” row of table 3 shows that the addition of the FE component resulted in a 0.17-SD improvement in child cognitive skills relative to HIM alone ($p = .005$).

This is a striking set of findings. On the one hand, we find no evidence that increasing per-child expenditure by nearly one-third had any impact on children’s cognitive development. On the other, the addition of the FE component, which cost a small fraction of the HIM component, resulted in sizeable, statistically significant impacts on cognitive development.

The last column of table 2 shows impacts on ASQ:SE, our measure of socioemotional problems scored according to the publisher’s guidelines; column 2 of table 3 presents impacts on the socioemotional problems factor constructed using the item responses to the ASQ:SE. Note that the ASQ:SE measures socioemotional problems, so higher values imply lower levels of socioemotional development. As with cognitive development, we find no evidence that the HIM program had any impact on socioemotional development. However, we also find no evidence that the HIM+FE program affected socioemotional development. Importantly, we may be underpowered to identify small impacts on this outcome—the larger standard errors in the socioemotional development analysis (table 3) indicate that these measures contain less information (see sec. IV.B).

B. Robustness

We test the robustness of these estimates to choice of scoring method, choice of control variables, and definition of analysis sample.

1. Scoring

It is reassuring that we see the same pattern of results using measures of cognitive and socioemotional development scored according to guidance by test publishers and those scored using a measurement model. In table B.7, we show further evidence that neither the magnitude nor the significance of our results is dependent on specific modeling choices we have made in estimating our factor scores. First, we show that our findings are robust to constructing child development factor scores without assuming

that the underlying latent factors are normally distributed in the control group. We use both nonparametric (cols. 1 and 8) and semiparametric methods (cols. 2 and 9) for estimating control group distributions and find very similar results. Second, as discussed in footnotes 11 and 12, we show that our results are robust to including guessing parameters for all items in the construction of the factors and to including items with very little variation (cols. 3 and 4). Third, we show that they are robust to relaxing the assumption of independence of errors across all items. In particular, we obtain very similar results if we adopt a nested structure where measurement errors may be correlated within a particular assessment but are independent across assessments (col. 5). Fourth, we show that we find similar results even if we do not impose that all cognitive assessments are measuring the same underlying factor although, in practice, an exploratory factor analysis does suggest a single underlying factor for all cognitive assessments (col. 6).¹⁸ Finally, we show that simply aggregating the scores calculated according to the guidelines produced by the test publishers through a linear factor model also gives very similar results (col. 7).

A concern in the construction and interpretation of our factor scores, which none of these checks address, is the possibility that treatment itself could alter the mapping between the underlying latent factors and item responses as discussed by Heckman, Pinto, and Savelyev (2013) and Heckman et al. (2020). Thus, we formally test the null hypothesis of measurement invariance for each item separately. In appendix D we show that the p -values from these tests are roughly uniformly distributed across the unit interval, just as we would expect if there were no underlying differences in the true measurement parameters across treatment groups. As an additional check, we construct factors that drop the small number of items where we reject the hypothesis of invariance and, separately, where we allow measurement parameters to vary by treatment status for these same items. Both approaches yield treatment effect estimates that are very similar to our main results (table D.1).

2. Controls

In table B.8, we show that our results are also not sensitive to the control variables we include. Even when we control for children's age alone, the p -values associated with the difference between HIM+FE and the control group and the difference in cognitive development between HIM+FE and HIM are, respectively, .073 and .016. Controlling for age and gender yields p -values of .048 and .017, respectively. Effect sizes and patterns of significance are left virtually unchanged with the addition of controls for city and baseline child development (either as raw scores or as factor scores).

¹⁸ The exploratory factor analysis yields a first factor with an eigenvalue of 2.99, while the second factor has an eigenvalue of just 0.28.

3. Sample

All of our main conclusions hold when we also include younger children (below 48 months at end line) for whom we have incomplete assessment data in our analysis; all significant impacts on individual child development assessments in table 2 remain statistically significant with similar estimated effect sizes (table B.9). Furthermore, in table B.10 we show that if we estimate child development factor scores in this full extended sample, including only the assessments that are available for all children in the extended sample, our estimates of the comparisons between HIM+FE and HIM and between HIM+FE and the pure control remain statistically significant at the 5% level and are not statistically different from estimated impacts using our main analysis sample. Finally, we test for heterogeneity by age within this extended sample and cannot reject the null that impacts are the same across the age distribution (table B.11).

C. Heterogeneity by Baseline Household Wealth

Several studies from high-income countries show that children from disadvantaged households benefit more from access to childcare than children from better-off backgrounds (Havnes and Mogstad 2015; Cornelissen et al. 2018; Felfe and Lalive 2018). Our results suggest that, conditional on being in childcare, more-disadvantaged children also benefit more from improvements in its quality. We capture household wealth using a wealth index constructed from data collected at baseline and define children from households that had an above-median wealth index as the “wealthier” group.¹⁹ Estimates in column 1 of table 4 show that neither the wealthier nor the more disadvantaged children experienced improvements in cognitive development as the result of the HIM program. In contrast, the HIM+FE program had a relatively large impact of 0.29 SD on cognitive development of children from poorer households and no impact on children from better-off households; the difference between the two groups is statistically significant. The impacts on socioemotional development are not significantly different from zero for either group (see col. 3).

D. Heterogeneity by Baseline Development

In line with the results above, we also find evidence of significant heterogeneity in impacts by level of child development at baseline. We define children with an above-median baseline development factor score, as

¹⁹ This wealth index was constructed by summarizing information about whether the household owned at least one of eight different assets (including, e.g., a car, a TV, a washing machine) through factor analysis.

TABLE 4
HETEROGENEITY BY WEALTH AND BASELINE CHILD DEVELOPMENT

	COGNITIVE DEVELOPMENT		SOCIOEMOTIONAL PROBLEMS	
	(1)	(2)	(3)	(4)
HIM	.028 (.112)	.103 (.117)	.007 (.210)	.095 (.204)
<i>p</i> -Value	.814	.374	.971	.645
HIM+FE	.286*** (.090)	.301*** (.092)	-.115 (.203)	-.146 (.210)
<i>p</i> -Value	.001	.002	.559	.486
HIM × wealthier	-.070 (.139)		-.046 (.199)	
<i>p</i> -Value	.615		.825	
HIM+FE × wealthier	-.252** (.125)		-.088 (.241)	
<i>p</i> -Value	.045		.716	
HIM × higher baseline development		-.227* (.131)		-.208 (.284)
<i>p</i> -Value		.081		.472
HIM+FE × higher baseline development		-.291** (.113)		-.015 (.266)
<i>p</i> -Value		.013		.964
Difference	.258*** (.076)	.197** (.087)	-.122 (.162)	-.241 (.163)
<i>p</i> -Value	.001	.018	.487	.136
Difference × wealthier	-.182* (.098)		-.042 (.241)	
<i>p</i> -Value	.063		.873	
Difference × higher baseline development		-.063 (.114)		.193 (.266)
<i>p</i> -Value		.578		.475
Observations	1,074	1,074	1,074	1,074

NOTE.—Two-sided *p*-values are estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) are in parentheses. We reestimate the measurement system on each bootstrapped sample. All estimates control for age, gender, city effects, and baseline scores for MacArthur-Bates CDI and each subscale of the ASQ-3 and ASQ:SE, as well as indicators of being above/below median on baseline wealth (cols. 1 and 3) and baseline child development (cols. 2 and 4). All factors are scaled so the underlying latent factor has a mean of 0 and SD of 1 in the control group. All factors are constructed as described in sec. IV. “Wealthier” implies that a child’s household had above-median value of household asset index at baseline. “Higher baseline development” implies that a child had above-median baseline child development as measured by the factor score discussed in sec. V.D.

* *p* < .1.
** *p* < .05.
*** *p* < .01.

measured by a factor aggregating the MacArthur-Bates CDI and the ASQ-3 administered at baseline (see sec. IV), as having “higher baseline development.”²⁰ We find that while HIM+FE resulted in a large and statistically

²⁰ When controlling for baseline child development in the main analysis, we included each of the subscales of ASQ-3 and MacArthur-Bates CDI separately. However, for this

significant improvement in cognitive skills of children with a lower-than-median level of development at baseline (0.30 SD), the program had no impact on children with higher-than-median baseline development (reported in col. 2 of table 4). The impacts on socioemotional development are not different from zero for either group (col. 4).

VI. Mechanisms

Why did increasing resources have no impacts on child development? Why did additionally providing training and information lead to positive impacts? In this section, we explore potential mechanisms, focusing on the responses of teachers to the programs, using data on the quality of the classroom learning environment and time use of teachers and TAs. We ground this analysis in a discussion of different margins on which teachers might respond to the provision of TAs and training, as well as the effects that economic theory would predict.

We focus on teachers' responses across three distinct margins. The first of these is the time and effort that teachers devote to classroom activities. Teachers in this setting had the scope to increase their overall classroom time input by working overtime to fit in administrative and preparation work. On the other hand, teachers who already routinely worked overtime had the scope to reduce their hours. The second margin is the division of teachers' time in the classroom between different types of activities, and the third is how teachers instructed any TAs they had to allocate TA classroom time. Teachers in HIs tend to have a high degree of autonomy over how they manage their class, so it is reasonable to assume that they chose how their time and that of their TAs was spent. The FE program emphasized the distinction between learning and caregiving activities, motivated by the concern that while the former are more important for child development, teachers tended to spend a lot of their time on the latter. We thus focus on understanding how teachers choose to divide classroom time between learning and caregiving activities.

The correlations in our data are consistent with the proposition that the choices that teachers make across these margins matter for child development. We utilize data on reported weekly hours of teacher overtime

heterogeneity analysis, we want to summarize all of these subscales into a single index. To do this, we age-standardize them by regressing our scores on dummies indicating a child's age in months and then residualizing. We then put all age-standardized measures into an exploratory factor analysis, which suggests that a single factor (with an eigenvalue of 1.7) meets the Kaiser criterion (Kaiser 1960). We predict this factor for all children and then divide children into those with below- and above-median baseline child development on the basis of this factor.

as a proxy for the total time that they spent on their job, data from the Teacher Survey of Early Education Quality to capture their allocation of time between learning and caring activities, and ECERS-R data to measure quality of teaching by teachers within the classroom (see sec. IV for details). In table 5, we report the results of a regression of the child cognitive-development factor (used in the main impact analysis in table 3) on our measures related to each of these potential mechanisms, averaged at the preschool level (e.g., average overtime of all teachers in the preschool). We include the same set of control variables as in the main impact analysis. We start by including each indicator individually. In line with the message of FE, columns 1 and 3 show a positive and significant association between child cognitive development and teacher-reported learning activities in the classroom as well as overtime, but no significant association with caring activities (col. 2). Column 4 further shows that good teaching processes that were directly observed during the ECERS-R assessment are positively correlated with children’s cognitive development. The magnitude of these correlations remains similar when we include indicators simultaneously (cols. 5–7) although the precision decreases in some cases. In the last column, we estimate the same regression as in column 7 but restrict the sample to children in the control

TABLE 5
CORRELATIONS BETWEEN CHILD COGNITIVE DEVELOPMENT
AND TEACHER-REPORTED ACTIVITIES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Learning activities	.084** (.033)				.097* (.050)	.090* (.047)	.078 (.056)	.100 (.101)
<i>p</i> -Value	.012				.054	.056	.167	.329
Caring activities		.021 (.030)			-.015 (.038)	-.017 (.038)	-.017 (.041)	-.035 (.082)
<i>p</i> -Value		.482			.703	.655	.689	.671
Overtime			.048* (.027)			.041 (.027)	.021 (.033)	.051 (.098)
<i>p</i> -Value			.081			.126	.524	.608
ECERS-R quality				.209* (.116)			.196* (.112)	.193 (.190)
<i>p</i> -Value				.076			.083	.318
Observations	1,074	1,074	1,074	726	1,074	1,074	726	249

NOTE.—Two-sided *p*-values and standard errors (in parentheses) are clustered at the HI level. Table presents ordinary least squares regression coefficients for regression of child cognitive-development factor on teachers’ involvement in learning and development activities, personal care activities, total overtime, and observed teaching quality as measured using the ECERS-R. Construction of all measures is described in sec IV. Routines, overtime, and ECERS-R quality measures are all averaged across all observations in the HI. All regressions control for city effects, child gender, child age, and baseline child development. Column 6 restricts the sample to children in the control group.

* *p* < .1.
** *p* < .05.

group only. The size of the coefficients does not change much, though the reduced sample size renders these estimates less precise.

Building on these insights, we turn to how we might expect teachers to respond to the two interventions and under what conditions. Here we provide an intuitive discussion that we formalize in an economic model presented in appendix E. For the reasons given above, we consider teachers' choices about how much time to spend in the classroom, as well as how to allocate this time between care and learning activities. We assume that, in making these choices, teachers care about both child development and leisure and that they make them subject to a time constraint and their beliefs about the child development production function. In this framework, the introduction of TAs can give rise to three distinct effects: a resource effect, a substitutability/complementarity effect, and a comparative advantage effect. The resource effect unambiguously reduces teachers' effort, in both learning and care activities; the new effort from TAs will improve child development for the same teacher effort inducing teachers to reallocate some of their time away from teaching and toward leisure. The direction of the substitutability/complementarity effect will depend on how (in the teachers' view) the addition of TAs alters the marginal product of teachers' effort. It will increase teachers' effort if teachers perceive their marginal product to be increasing in the addition of TAs and decrease their effort if they perceive the reverse as true. Finally, comparative advantage will guide how teachers reallocate their time between different activities. They will spend more time on activities in which their time is more complementary with that of the TAs and less on those in which it is more substitutable.

The total time teachers spend on classroom activities will depend on only the resource and substitutability/complementarity effects. If teachers believe that the addition of TAs reduces the marginal product of their own time, then our framework suggests that teachers will respond by unambiguously reducing their total teaching efforts. If, on the other hand, teachers perceive that TAs increase the marginal productivity of teachers, then adding TAs will have an ambiguous effect on teachers' time. If the resource effect dominates the complementarity effect, teachers will reduce the overall time they put in, whereas if the complementarity effect dominates, they will increase the overall time they put in.

Table 6 presents estimated treatment effects of HIM and HIM+FE on overtime done by the teachers (col. 1), the frequency with which learning and care activities are undertaken by the teachers (cols. 2 and 3) and the TAs (cols. 4 and 5), and, finally, quality of teaching delivered by the teachers (col. 6).

We find that teachers responded to the HIM program by reducing the frequency with which they performed both learning and personal care activities in the classroom, as well as the amount of overtime they worked.

TABLE 6
IMPACTS ON TEACHERS' AND TAs' BEHAVIOR

	TEACHERS' OVERTIME	TEACHERS' ACTIVITIES		TA's ACTIVITIES		OBSERVED TEACHING QUALITY
		Learning	Care	Learning	Care	
	(1)	(2)	(3)	(4)	(5)	(6)
HIM only	-.450** (.218)	-.359** (.171)	-.020*** (.270)			.008 (.095)
<i>p</i> -Value	.035	.038	.002			.936
FE+HIM	-.033 (.275)	-.045 (.162)	-.758** (.270)	.186 (.217)	.019 (.248)	.180** (.093)
<i>p</i> -Value	.911	.772	.012	.398	.923	.050
Difference	.417 (.270)	.314** (.152)	.262 (.241)			.172* (.105)
<i>p</i> -Value	.129	.040	.268			.089
Observations	841	841	839	254	254	172

NOTE.—Two-sided *p*-values are estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) are in parentheses. We reestimate the measurement system on each bootstrapped sample. All estimates control for HI-level averages of baseline measures of the outcome in question. Observed teaching quality controls for the average age of kids in the classroom. All regressions control for city effects other than in col. 6, for which such regressions are not possible because we lack the ECERS-R measures for all HIs and one city does not contain all treatment groups. Overtime is measured in hours per week. The other variables are factor scores scaled so the underlying latent factor has a mean of 0 and SD of 1 in the control group (HIM group in the case of TA activities). All factors are constructed as described in sec. IV.C.

* *p* < .1.
** *p* < .05.
*** *p* < .01.

The impacts are significant and sizeable: there was a 0.36 SD reduction in frequency of learning activities and a 1.0 SD reduction in the frequency of personal care activities. Teacher-reported overtime also fell by nearly half an hour per week (relative to 1.2 hours among teachers in the control group). Furthermore, conditional on being present in the classroom, we see no evidence of a change in the quality of teaching delivered by the teachers; HIM has no impact on psychologists' assessment of the overall quality of teaching by the teachers during classroom observation.²¹ Overall, while HIM appears to have had no effect on the quality of teachers' teaching conditional on being present, we see clear evidence that HIM led teachers to scale back the time they spend on all classroom activities. Drawing on the discussion above, this reduction is consistent with a high marginal valuation of leisure among teachers (leading to a strong resource effect) and/or teachers holding the belief that TAs are highly substitutable with their own efforts.

²¹ Note that we have these measures for only 172 out of the 841 classrooms in the sample (see sec. IV.C for details).

Teachers' choices may change when the TA program is paired with training, as in HIM+FE, if this training changes teachers' beliefs about the child development production function. Such a change could reflect a change in true productivity if the training made teachers more productive at some or all classroom activities. For instance, the training included practical advice on how to plan and implement high-quality learning activities. In addition, the information about the process of child development that the training delivered could have changed teachers' perceptions of the process and, as a consequence, the emphasis teachers placed on different activities even without changing their productivity in performing any given activity. For example, if teachers previously underestimated the productivity of learning activities, they may have dedicated less of their time to these activities than they would under full information. We formalize this intuition in appendix E.

Table 6 shows that the addition of the FE training did indeed change teacher behavior. Specifically, relative to receiving HIM alone, the addition of FE offset the reduction in the time teachers dedicate to learning activities, as well as the reduction in overtime, but not the reduction in time spent on care activities.²² Thus, while we observe a reduction in personal care activities relative to the control group of roughly the same size in the HIM and HIM+FE arms, the negative effects that we see on learning activities or overtime in the HIM arm are not there in the HIM+FE arm.

Our results suggest that, on the margin, the addition of FE led to an increase in how useful teachers consider time devoted to learning activities to be for child development relative to that devoted to care activities. In line with the discussion above, there are several compatible channels that might underlie such a shift. First, FE training may have resulted in a real increase in teachers' productivity in running learning activities prompting them to devote a greater portion of their time to these activities. Second, even without generating changes in true productivity, FE may have corrected misperceptions teachers held about the child development production function. Specifically, our results are consistent with teachers revising upwards their perception of how useful time spent on learning activities is for promoting child development and thus reallocating

²² Our results suggest that FE increased the amount of effort that teachers in a given HI allocated to classroom teaching and to learning activities in particular. It is most natural to think that FE increased the effort of the teachers who were already employed by the HI. FE could also have resulted in HIs hiring teachers who exerted a higher level of effort. We consider this to be less plausible both because it is conceptually unclear how training existing teachers would change hiring practices and because we see the exact same pattern of impacts on teacher behavior if we restrict the sample to teachers who were employed in the center at baseline (see table B.12).

their time toward these activities.²³ Both of these channels point to an increase in the overall productivity of teachers' time. We see this reflected in column 6. Even with the small sample size for this measure, we see evidence that the addition of FE was effective at improving the quality of teaching (as observed by psychologists). Compared to the pure control, we estimate that the HIM+FE program improved the quality of directly observed teaching by 0.18 SD ($p = .050$), with the difference between the HIM and HIM+FE arms being similar in magnitude and statistical significance.²⁴

To sum up, these results suggest that teachers' behavioral reactions are key to understanding both the null effects of HIM and the positive effects of HIM+FE on child development. They are consistent with the idea that, in the HIM arm, teachers used TAs to substitute their time in all activities, irrespective of the importance of different activities for child development or the training and experience needed to execute them well. This could explain why we see no improvements in child development. The training delivered through FE, however, may have provided teachers with a better understanding of the process of child development and productive teaching approaches, enabling them to integrate the TAs into the classroom and/or adapt their own activities in the classroom in a way that was conducive to improvements in children's development.

VII. Conclusions

In this paper, we have shown that, even within the same institutional setting, different approaches to improving the quality of early-years education can have very different effects on child development. We present

²³ In app. E, we outline how such a shift could be driven either by FE increasing teachers' assessment of the overall usefulness of learning activities or by increasing teachers' perception of their own comparative advantage in learning activities relative to their TAs. Formally, these two channels would lead to opposite predictions for TA time use: the first would suggest that TAs would also increase their effort in learning activities while the second would suggest that they would exploit their comparative advantage in caring activities. Thus, the fact that we find no significant impacts on TA time use (cols. 4 and 5) could be suggestive of both mechanisms being at play and thus offsetting each other. However, we note that our estimates here are imprecise, and so we do not draw strong conclusions.

²⁴ If we were to go further and assume that the quality of teaching observed during the ECERS-R assessment summarized all aspects of classroom quality relevant to child development and was the only channel through which the treatments affects cognitive development, we could instrument for observed quality using treatment assignment to obtain estimates of the local average treatment effect. We show the results of this exercise in table B.15. Reassuringly, they are very similar to the ratio of the "reduced-form" and "first-stage" coefficients shown in tables 3 and 6 respectively and suggest that a 1-SD improvement in observed teaching quality in a preschool translated into a 1.24-SD improvement in cognitive development. We note, however, that the above assumptions are very strong. In particular, because teachers were always present in the classroom while the observation was happening, ECERS-R cannot capture any impacts on teaching quality that come from teachers altering the amount of time that they spend in the classroom, which we show to be important in table 6.

the striking finding that a costly national government program that provided preschools with resources to hire TAs had no impact on child development. In contrast, also including—at little extra cost—a professional development training program for existing preschool teachers resulted in significant positive overall impacts on children's cognitive development of around 0.16 SD of the control group and especially large benefits of 0.29 SD for the more disadvantaged children in the sample.

These are sizable impacts. To the extent that credible comparisons can be made between studies, 0.17 SD corresponds to 23% of the achievement gap between children in the top and bottom wealth quintiles in Colombia at age 6 (Rubio-Codina and Grantham-McGregor 2019) and is in the ballpark of studies that evaluate effects of children, on the extensive margin, accessing center-based care in Colombia (Nores, Bernal, and Barnett 2019) and other Latin American countries (Berlinski, Galiani, and Manacorda 2008; Noboa-Hidalgo and Urzúa 2012; Bernal and Fernández 2013; Behrman et al. 2014; Bernal and Ramírez 2019). 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 United States, suggests that programs that achieved short-run effects of similar magnitude can have wide-ranging and persistent positive long-run effects (Garces, Thomas, and Currie 2002; Deming 2009).

We provide some insights into the mechanisms driving the starkly different impacts that we find for the two interventions. We show that provision of TAs resulted in teachers reducing their time at work, including on learning activities, that are positively correlated with child development. The addition of the teacher training program, however, induced teachers to increase time spent at work, including on learning activities. The null effect of additional TAs may have been generated by teachers placing a high marginal value on reallocating some of their overtime into leisure. Such an effect would have been more likely if teachers believed that their role in the classroom is highly substitutable with that of the TAs. The teacher time-use response to the FE teacher professional development training program is consistent with a change in their beliefs about the relative importance of different activities in the classroom for child development. This could have been due to actual changes in how effective teachers are at performing various activities or due to a correction of misperceptions teachers held about the role of different inputs in the process of child development.

Our findings complement a recent set of studies showing that more intensive use of unskilled teachers and TAs can be effective at improving learning outcomes, as discussed by Banerjee et al. (2017) in relation to the successful scale-up of Teaching at the Right Level in India and by Duflo, Kiessel, and Lucas (2020) when describing interventions that introduced

TAs to primary schools in Ghana. Of course, these studies span very different contexts, so findings of differential effectiveness of similar interventions is not surprising. However, it is also plausible that these studies and our findings are telling a similar story. Most of the interventions analyzed in these studies provided not only TAs but also a clear set of tasks for these TAs to undertake, which was not the case in the HIM intervention. The addition of the teacher professional development training, by contrast, may have given teachers the skills needed to delegate tasks to TAs appropriately. This evidence suggests that in contexts where teachers are poorly trained, additional school resources can be effective when accompanied by guidance on how to utilize them. Without guidance, however, such provision might generate unintended and undesirable consequences, such as the reduction in effort that we see among teachers in the HIM program.

Data Availability

Code replicating the tables and figures in this article can be found in Andrew et al. (2023) in the Harvard Dataverse, <https://doi.org/10.7910/DVN/GORPVP>.

References

- Agostinelli, Francesco, Ciro Avitabile, and Matteo Bobba. 2023. "Enhancing Human Capital in Children: A Case Study on Scaling." Working Paper no. 31407 (June), NBER, Cambridge, MA.
- Agostinelli, Francesco, and Matthew Wiswall. 2016. "Estimating the Technology of Children's Skill Formation." Working Paper no. 22442 (July), NBER, Cambridge, MA.
- Andrew, Alison, Orazio Attanasio, Raquel Bernal, Lina Cardona Sosa, Sonya Krutikova, and Marta Rubio-Codina. 2023. "Replication Data for: 'Preschool Quality and Child Development.'" Harvard Dataverse. <https://doi.org/10.7910/DVN/GORPVP>.
- Andrew, Alison, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2018. "Impacts 2 Years after a Scalable Early Childhood Development Intervention to Increase Psychosocial Stimulation in the Home: A Follow-Up of a Cluster Randomised Controlled Trial in Colombia." *PLoS Medicine* 15 (4): e1002556.
- Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Q.J.E.* 131 (3): 1415–53.
- Araujo, M. Caridad, and Norbert Schady. 2015. "Daycare Services: It's All about Quality." In *The Early Years: Child Well-Being and the Role of Public Policy*, edited by Samuel Berlinski and Norbert Schady, 91–121. New York: Palgrave Macmillan.
- Attanasio, Orazio, Sarah Cattán, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020. "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia." *A.E.R.* 110 (1): 48–85.
- Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020. "Human Capital Development and Parental Investment in India." *Rev. Econ. Studies* 87 (6): 1125–41.

- Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *J. Econ. Perspectives* 31 (4): 73–102.
- Banerjee, A., S. Cole, E. Duflo, and L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Q.J.E.* 122 (3): 1235–64.
- Bartlett, Maurice S. 1937. "The Statistical Conception of Mental Factors." *British J. Psychology* 28 (1): 97–104.
- Behrman, Jere R., John Hoddinott, John A. Maluccio, et al. 2014. "What Determines Adult Cognitive Skills? Influences of Pre-School, School, and Post-School Experiences in Guatemala." *Latin American Econ. Rev.* 23:4.
- Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. "The Effect of Pre-Primary Education on Primary School Performance." *J. Public Econ.* 93 (1–2): 219–34.
- Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. "Giving Children a Better Start: Preschool Attendance and School-Age Profiles." *J. Public Econ.* 92 (5–6): 1416–40.
- Bernal, Raquel, Orazio Attanasio, Ximena Peña, and Marcos 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 Res. Q.* 47 (2): 418–31.
- Bernal, Raquel, and Camila Fernández. 2013. "Subsidized Childcare and Child Development in Colombia: Effects of *Hogares Comunitarios de Bienestar* as a Function of Timing and Length of Exposure." *Soc. Sci. and Medicine* 97:241–49.
- Bernal, Raquel, and Sara María Ramírez. 2019. "Improving the Quality of Early Childhood Care at Scale: The Effects of "From Zero to Forever." *World Development* 118:91–105.
- Birnbaum, Allan. 1968. "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In *Statistical Theories of Mental Test Scores*, edited by Frederic M. Lord and Melvin R. Novick, 397–479. Reading, MA: Addison-Wesley.
- Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng, and Justin Sandefur. 2018. "Experimental Evidence on Scaling Up Education Reforms in Kenya." *J. Public Econ.* 168:1–20.
- Britto, Pia Rebello, Hirokazu Yoshikawa, and Kimberly Boller. 2011. "Quality of Early Childhood Development Programs in Global Contexts: Rationale for Investment, Conceptual Framework and Implications for Equity." *Soc. Policy Report* 25:2.
- Burchinal, Margaret R., Joanne E. Roberts, Rhodus Riggins Jr., Susan A. Zeisel, Eloise Neebe, and Donna Bryant. 2000. "Relating Quality of Center-Based Child Care to Early Cognitive and Language Development Longitudinally." *Child Development* 71 (2): 39–357.
- Caucutt, Elizabeth M., Lance Lochner, and Youngmin Park. 2017. "Correlation, Consumption, Confusion, or Constraints: Why Do Poor Children Perform So Poorly?" *Scandinavian J. Econ.* 119 (1): 102–47.
- Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Q.J.E.* 126 (4): 1593–660.
- Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2018. "Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance." *J.P.E.* 126 (6): 2356–409.

- Cunha, Flávio, Irma Elo, and Jennifer Culhane. 2022. "Maternal Subjective Expectations about the Technology of Skill Formation Predict Investments in Children One Year Later." *J. Econometrics* 231:3–32.
- Cunha, Flávio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931.
- Danzer, Virginia A., Mary Frances Gerber, Theresa M. Lyons, and Judith K. Voress. 1991. *Daberon 2: Screening for School Readiness*. Austin, TX: Pro-Ed.
- Das, Jishnu, and Tristan Zajonc. 2010. "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement." *J. Development Econ.* 92 (2): 175–87.
- Datta Gupta, Nabanita, and Marianne Simonsen. 2010. "Non-Cognitive Child Outcomes and Universal High Quality Child Care." *J. Public Econ.* 94 (1–2): 30–43.
- Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *J. Econ. Literature* 48 (2): 424–55.
- de Chaisemartin, Clément, and Jaime Ramirez-Cuellar. 2020. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" Working Paper no. 27609 (July), NBER, Cambridge, MA.
- Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Econ. J. Appl. Econ.* 1 (3): 111–34.
- Diamond, A., and C. Taylor. 1996. "Development of an Aspect of Executive Control: Development of the Abilities to Remember What I Said and to 'Do as I Say, Not as I Do.'" *Developmental Psychobiology* 29 (4): 315–34.
- Duflo, Annie, Jessica Kiessel, and Adrienne M. Lucas. 2020. "External Validity: Four Models of Improving Student Achievement." Working Paper no. 27298 (June), NBER, Cambridge, MA.
- Dunn, Lloyd M., Eligio R. Padilla, Delia E. Lugo, and Leota M. Dunn. 1986. *Test de Vocabulario en Imágenes Peabody (TVIP)*. Circle Pines, MN: AGS.
- Elango, Sneha, Jorge Luis García, James J. Heckman, and Andrés Hojman. 2015. "Early Childhood Education." In *Economics of Means-Tested Transfer Programs in the United States*, vol. 2, edited by Robert A. Moffitt, 235–97. Chicago: Univ. Chicago Press (for NBER).
- Engle, Patrice L., Lia C. H. Fernald, Harold Alderman, et al. 2011. "Strategies for Reducing Inequalities and Improving Developmental Outcomes for Young Children in Low-Income and Middle-Income Countries." *Lancet* 378 (9799): 1339–53.
- Evans, David K., and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Res. Observer* 31 (2): 242–70.
- Felfe, Christina, and Rafael Lalive. 2018. "Does Early Child Care Affect Children's Development?" *J. Public Econ.* 159 (January): 33–53.
- Fort, Margherita, Andrea Ichino, and Giulio Zanella. 2020. "Cognitive and Non-cognitive Costs of Day Care at Age 0–2 for Children in Advantaged Families." *J.P.E.* 128 (1): 158–205.
- Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *A.E.R.* 92 (4): 999–1012.
- Gerber, Susan B., Jeremy D. Finn, Charles M. Achilles, and Jayne Boyd-Zaharias. 2001. "Teacher Aides and Students' Academic Achievement." *Educ. Evaluation and Policy Analysis* 23 (2): 123–43.

- Glewwe, Paul, Eric Hanushek, Sarah Humpage, and Renato Ravina. 2011. "School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010." Working Paper no. 17554 (October), NBER, Cambridge, MA.
- Glewwe, P. and K. Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In *Handbook of the Economics of Education*, vol. 5, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 653–743. Amsterdam: North-Holland.
- Hallam, R., B. Rous, S. Riley-Ayers, and D. Epstein. 2011. *Teacher Survey of Early Education Quality*. New Brunswick, NJ: Nat. Inst. Early Educ. Res.
- Hanushek, Eric A. 1999. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." *Educ. Evaluation and Policy Analysis* 21 (2): 143–63.
- Harms, T., R. M. Clifford, and D. Cryer. 1998. *Early Childhood Environment Rating Scale*. New York: Teachers Coll. Press.
- Havnes, Tarjei, and Magne Mogstad. 2015. "Is Universal Child Care Leveling the Playing Field?" *J. Public Econ.* 127:100–114.
- Heckman, James J. 1992. "Randomization and Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, edited by Charles Manski and Irwin Garfinkel, 201–30. Cambridge, MA: Harvard Univ. Press.
- Heckman, James J., Bei Liu, Mai Lu, and Jin Zhou. 2020. "The Impacts of a Prototypical Home Visiting Program on Child Skills." Working Paper no. 27356 (June), NBER, Cambridge, MA.
- Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the HighScope Perry Preschool Program." *J. Public Econ.* 94 (1–2): 114–28.
- Heckman, James J., and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Ann. Rev. Econ.* 6:689–733.
- Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms through Which an Influential Early Childhood Program Boosted Adult Outcomes." *A.E.R.* 103 (6): 2052–86.
- Heckman, James J., and Jin Zhou. 2022. "Measuring Knowledge and Learning." Working Paper no. 29990 (April), NBER, Cambridge, MA.
- Jackson-Maldonado, Donna, Virginia A. Marchman, and Lia C. H. Fernald. 2013. "Short-Form Versions of the Spanish MacArthur-Bates Communicative Development Inventories." *Appl. Psycholinguistics* 34 (04): 837–68.
- Jackson-Maldonado, Donna, Donna J. Thal, Larry Fenson, Virginia A. Marchman, Tyler Newton, Barbara T. Conboy, and Elizabeth Bates. 2003. *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's Guide and Technical Manual*. Baltimore: Brookes.
- Joo, Young Sun, Katherine Magnuson, Greg J. Duncan, Holly S. Schindler, Hirokazu Yoshikawa, and Kathleen M. Ziol-Guest. 2020. "What Works in Early Childhood Education Programs?: A Meta-analysis of Preschool Enhancement Programs." *Early Educ. and Development* 31 (1): 1–26.
- Kaiser, Henry F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educ. and Psychological Measurement* 20 (1): 141–51.
- Kline, Patrick, and Christopher R. Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *Q.J.E.* 131 (4): 1795–848.
- Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Q.J.E.* 114 (2): 497–532.

- Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in the Early Grades on College Test Taking and Middle School Test Results: Evidence from Project STAR." *Econ. J.* 111 (468): 1–28.
- List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple Hypothesis Testing in Experimental Economics." *Experimental Econ.* 22:773–93.
- Neuman, Michelle J., and Lynette Okeng'o. 2019. "Early Childhood Policies in Low- and Middle-Income Countries." *Early Years* 39:223–28.
- Noboa-Hidalgo, Grace E., and Sergio S. Urzúa. 2012. "The Effects of Participation in Public Child Care Centers: Evidence from Chile." *J. Human Capital* 6 (1): 1–34.
- Nores, Milagros, Raquel Bernal, and W. Steven Barnett. 2019. "Center-Based Care for Infants and Toddlers: The aeioTU Randomized Trial." *Econ. Educ. Rev.* 72:30–43.
- Özler, Berk, Lia C. H. Fernald, Patricia Kariger, Christin McConnell, Michelle Neuman, and Eduardo Fraga. 2018. "Combining Pre-School Teacher Training with Parenting Education: A Cluster-Randomized Controlled Trial." *J. Development Econ.* 133:448–67.
- Peisner-Feinberg, Ellen S., Margaret R. Burchinal, Richard M. Clifford, et al. 2001. "The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Developmental Trajectories through Second Grade." *Child Development* 72 (5): 1534–53.
- Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning*. Baltimore: Brookings Inst. Press (for Center for Global Development).
- Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4): 1237–82.
- . 2010. "Balanced Control of Generalized Error Rates." *Ann. Statist.* 38 (1): 598–633.
- Rosero, José, and Hessel Oosterbeek. 2011. "Trade-Offs between Different Early Childhood Interventions: Evidence from Ecuador." Discussion Paper no. TI 2011-102/3, Tinbergen Inst., Amsterdam.
- Rubio-Codina, Marta, and Sally Grantham-McGregor. 2019. "Evolution of the Wealth Gap in Child Development and Mediating Pathways: Evidence from a Longitudinal Study in Bogota, Colombia." *Developmental Sci.* 22 (5): e12810.
- Schrank, Fredrick A., Kevin S. McGrew, Mary L. Ruef, Criselda G. Alvarado, Ana F. Muñoz-Sandoval, and Richard W. Woodcock. 2005. *Overview and Technical Supplement (Batería III Woodcock-Muñoz Assessment Service Bulletin No. 1)*. Itasca, IL: Riverside.
- Singh, Abhijeet. 2020. "Learning More with Every Year: School Year Productivity and International Learning Divergence." *J. European Econ. Assoc.* 18 (4): 1770–813.
- Squires, Jane, Diane Bricker, and Elizabeth Twombly. 2002. *Ages & Stages Questionnaires: Social-Emotional*. Baltimore: Brookes.
- . 2009. *Ages & Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed Child-Monitoring System*. Baltimore: Brookes.
- Sylva, Kathy, Iram Siraj-Blatchford, Brenda Taggart, Pam Sammons, Eduard Melhuish, Karen Eliot, and Vasiliki Totsika. 2006. "Capturing Quality in Early Childhood through Environmental Rating Scales." *Early Childhood Res. Q.* 21 (1): 76–92.
- Van Der Linden, Wim J., and Ronald K. Hambleton. 1997. *Handbook of Modern Item Response Theory*. New York: Springer.

- Wolf, Sharon. 2018. "Impacts of Pre-Service Training and Coaching on Kindergarten Quality and Student Learning Outcomes in Ghana." *Studies in Educ. Evaluation* 59:112–23.
- Woodcock, Richard W. 1977. "Woodcock-Johnson Psycho-Educational Battery." Technical report, Teaching Resources, Boston.
- World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise..* Washington, DC: World Bank.
- Yoshikawa, Hirokazu, Diana Leyva, Catherine E. Snow, Ernesto Treviño, M. Clara Barata, Christina Weiland, Celia J. Gomez, et al. 2015. "Experimental Impacts of a Teacher Professional Development Program in Chile on Preschool Classroom Quality and Child Outcomes." *Developmental Psychology* 51 (3): 309–22.
- Yoshikawa, Hirokazu, Alice J. Wuermli, Abbie Raikes, Sharon Kim, and Sarah B. Kabay. 2018. "Toward High-Quality Early Childhood Development Programs and Policies at National Scale: Directions for Research in Global Contexts." *Soc. Policy Report* 31 (1): 1–36.