

Preschool Quality and Child Development^{*†}

Alison Andrew[‡] Orazio P. Attanasio[§] Raquel Bernal[¶] Lina Cardona Sosa^{||}

Sonya Krutikova,^{**} Marta Rubio-Codina^{††}

March 28, 2023

Abstract

Globally, access to preschool has increased dramatically but its quality is often poor. We evaluate two interventions aimed at improving the quality of public preschools in Colombia. The first, designed by the government and rolled-out nationwide, provided preschools with significant extra funding, mainly earmarked for hiring teaching assistants (TAs). The second, for a small additional cost, also offered training for existing teachers. We show that the first intervention did not improve child development, while the second led to significant improvements in children's cognitive development, especially for those from more disadvantaged backgrounds. We argue these dramatic differences can be explained by the two interventions having different impacts on teachers' behavior. The first led teachers to reduce the time they spent in the classroom, including on learning activities. The addition of the training offset this adverse effect of TA provision on teachers' learning activities and improved the quality of teaching.

*We thank Diana Pérez López and Diana Martínez Heredia for excellent research assistance and gratefully acknowledge the contributions of Carlos Medina and Marcos Vera-Hernández to the design of this study and of Ximena Peña to both study design and implementation. Ximena passed away in January 2017 and is dearly missed. We thank James Heckman and three anonymous referees for extremely useful comments and suggestions. We thank the Jacobs Foundation for hosting us at the Marbach Residence Program in 2017 where we made significant progress on this project.

[†]This research was funded by the International Initiative for Impact Evaluation (3ie) and Fundación Éxito. Prof. Attanasio acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 695300-HKADeC-ERC-2015-AdG). Andrew and Krutikova acknowledge funding from the ESRC Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. Prof. Bernal acknowledges funding from the British Academy Visiting Fellowship VF1 10124. The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing up results. Ethics Committees at Universidad de los Andes and University College London approved the study's protocol in 2013.

[‡]Oxford University & IFS

[§]Yale & IFS

[¶]Universidad de los Andes

^{||}World Bank

^{**}Manchester University & IFS

^{††}IFS & InterAmerican Development Bank

1 Introduction

It is now widely accepted that well-designed early childhood education 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 pre-primary education is rising fast; it increased from 29% in 1990 to 49% in 2015.¹ However, as governments expand coverage of early childhood education (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 (Rosero and Oosterbeek, 2011; Engle et al., 2011; Britto, Yoshikawa, and Boller, 2011; Araujo and Schady, 2015; Ichino, Fort, and Zanella, 2019).

This issue is particularly relevant for lower- and middle-income countries (LMICs) where, according to the (limited) available evidence, ECE services are of very varied quality with many children receiving poor-quality center-based care (Araujo and Schady, 2015; Yoshikawa et al., 2018). Many LMICs are resorting to adding pre-primary classes to existing primary schools without allocating sufficient resources or expertise to ensure 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 homecare; 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 US) 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” in line with the acronym 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 was designed to complement the first by additionally providing professional development training

¹Figures from World Bank EdStats’ ‘Gross enrolment ratio, pre-primary, both sexes (%)’ series, available from <https://data.worldbank.org/data-catalog/ed-stats>. This definition gives the total enrollment in pre-primary education, regardless of age, as a percentage of pre-primary-age population. It classifies pre-primary education as ‘Education designed to support early development in preparation for participation in school and society. Programmes designed for children from age 3 to the start of primary education’.

for existing preschool teachers. We label it “HIM+FE”, again aligning with the government acronym.

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 SDs. 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 thirty percent of the control group standard deviation.

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 upwards or downwards. Teachers also have a high degree of autonomy over how they manage their class and can adjust how they split their time between different teaching activities, as well as 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 which 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 be suggestive of them not fully exploiting what we presume to be their 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 the FE program 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 how useful they thought spending time in this way was for child development (Caucutt, Lochner, and Park, 2017; Cunha, Elo, and Culhane, 2022).

Taken together, the impacts we find on child development and teachers 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 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 early childhood education 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; Heckman, Pinto, and Savelyev, 2013; Engle et al., 2011). 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 these estimates bypass the frequent uncertainties about whether program impacts estimated through 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 early childhood education 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 STAR 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 (2021) 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 not having clearly defined tasks and teachers having scope to endogenously react to the increase in TAs by reducing their own effort may have contributed to explaining 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 (Özler et al., 2018; Yoshikawa et al., 2015). 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 impact 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 2 provides details about the study setting and the interventions that we evaluate. Section 3 presents the study design and empirical strategy we use. In Section 4, we describe our outcome measures and how we use them in the analysis. The estimates of the main impacts are presented in Section 5, alongside robustness checks and heterogeneity analysis. In Section 6, we explore potential mechanisms using novel data on teacher time-use and quality of the classroom learning environment. Section 7 concludes.

2 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. This is the oldest public center-based childcare provider in Colombia and has enrolled an average of 125,000 children per year over the last 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 three and ten 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, to 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 one and a half 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 S\$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 *Hogares Infantiles* was announced and the new intervention was labeled *Hogares Infantiles Mejorados* (“Improved HI”; HIM). Specifically, HIs were given a substantial amount of additional resources, mainly for hiring new staff. The

²Occasionally, HIs take children as young as 6 months when it is “proven that they do not have a responsible adult to care for them”. However, the vast majority of children enrolled in HIs are 18 months or older.

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 a randomized controlled trial (RCT) into the initial HIM rollout. To this end, a random subset of HIs were wait-listed to receive the program a 18 months later. Additionally, there was interest from a well-established Colombian NGO, *Fundación Éxito* (FE), 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 hiring resources through HIM *and* teacher professional development training through FE. The training program was developed by FE 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 video-conferencing 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 due to 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

³This paper focuses on impacts on child development. In Appendix 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.

⁴We find no impact on any indicator of reading routines in the home. See Appendix 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 Appendix 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.

relative to the “business-as-usual” unenhanced model, which amounted to extra expenditure of \$300 per 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 upfront 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 for details of calculations.

3 Study Design and Empirical Strategy

We designed a three-armed cluster randomized controlled trial around the national rollout of the HIM program in order to assess effects of HIM alone and 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. 40 HIs were randomized into each of the three arms: (i) HIM, where preschools received the government quality improvement program; (ii) HIM+FE, where preschools received the teacher professional development training enhancement in addition to the HIM program; and (iii) a pure control group where the implementation of HIM was delayed. This design allows us to test 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 GPS 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 to 36 months at baseline). Within each triplet of eligible HIs, we randomly assigned one HI to the pure control group, one HI to the HIM treatment group and one HI to the HIM+FE treatment group. Randomization and sample selection were carried out over November–December 2012.

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 to 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 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). Endline 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

⁵We 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.

age of 48 months. Therefore, our main analysis sample comprises only children above 48 months at endline who were thus eligible for all assessments. We return to this issue in Section 5.2

3.1 Balance and Attrition

Attrition was relatively low. We completed *some* endline child development assessments for all but 155 children (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 1074 children with complete assessment data in our main analysis sample. The attrition rate amongst 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 socio-demographic 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 socio-emotional skills. We do see slight imbalances in fine motor 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

The majority of children (72.2%) continued attending the same HI throughout the study period; by endline, 9.2% were enrolled in a different HI (mostly one not in the study sample), 13.1% were enrolled in a different public or private 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 impacted 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.

3.2 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 data we have. 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 teachers, TAs, nutritionists and socioemotional experts, collected at baseline and endline to calculate what proportion of the budget allocated for hiring the

⁶In our robustness analysis in Section 5.2 we show we obtain similar results when examining either sample.

⁷When 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 endline cognitive and socioemotional development. In contrast, baseline measures of language, problem solving and communication skills are highly predictive of endline 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	N
Male	0.456 (0.499)	0.552 (0.498)	0.522 (0.500)	<i>p</i> =0.007	<i>p</i> =0.079	<i>p</i> =0.468	1074
Age (months)	32.98 (2.120)	32.77 (2.179)	32.73 (2.200)	<i>p</i> =0.287	<i>p</i> =0.166	<i>p</i> =0.836	1074
HH income (million COP)	1333.1 (774.2)	1341.5 (777.6)	1337.5 (795.5)	<i>p</i> =0.901	<i>p</i> =0.959	<i>p</i> =0.957	1074
Mother's education (years)	12.63 (2.776)	12.37 (2.601)	12.66 (2.579)	<i>p</i> =0.301	<i>p</i> =0.864	<i>p</i> =0.131	1064
Father's education (years)	12.01 (3.041)	11.98 (3.116)	12.13 (3.073)	<i>p</i> =0.911	<i>p</i> =0.699	<i>p</i> =0.541	1003
Household size	3.385 (1.697)	3.477 (1.629)	3.217 (1.542)	<i>p</i> =0.517	<i>p</i> =0.170	<i>p</i> =0.061	1074
ASQ Communication	63.95 (19.77)	65.86 (20.84)	64.33 (20.13)	<i>p</i> =0.326	<i>p</i> =0.843	<i>p</i> =0.424	1074
ASQ Gross Motor	62.22 (21.67)	66.22 (20.89)	64.53 (20.03)	<i>p</i> =0.059	<i>p</i> =0.178	<i>p</i> =0.417	1074
ASQ Problem Solving	57.63 (19.51)	59.40 (20.35)	58.71 (19.20)	<i>p</i> =0.414	<i>p</i> =0.589	<i>p</i> =0.721	1074
ASQ Personal Social	57.86 (18.59)	60.49 (18.59)	59.01 (18.37)	<i>p</i> =0.151	<i>p</i> =0.508	<i>p</i> =0.330	1074
ASQ Fine Motor	46.98 (20.09)	51.50 (20.81)	46.56 (19.73)	<i>p</i> =0.070	<i>p</i> =0.875	<i>p</i> =0.037	1074
MacArthur-Bates Language	66.16 (24.09)	67.68 (24.03)	66.49 (23.41)	<i>p</i> =0.566	<i>p</i> =0.911	<i>p</i> =0.647	1074
ASQ Socio-Emotional	56.09 (21.42)	53.29 (19.73)	54.61 (20.65)	<i>p</i> =0.137	<i>p</i> =0.490	<i>p</i> =0.466	1074
N	353	384	337				

Note. Baseline means (SDs) by treatment status for children included in the analysis sample (i.e. all children with complete child development assessment data at endline). Two-sided *p*-values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). ASQ child development scores are the raw scores from the five subscales of the ASQ: 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 Section 4.2.

additional personnel was spent by HIs in this way. 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 endline, 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 suggests 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 to 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 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, including TAs, directors or other senior staff, also participated. We do not have information on numbers or characteristics of teachers who were nominated by the center director and which of these 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.

3.3 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 ($T_{lm}^{HIM} = 1$) or HIM+FE ($T_{lm}^{HIM+FE} = 1$) on final outcomes through ordinary least squares (OLS):

$$Y_{ilm} = \beta_0 + \beta_1 T_{lm}^{HIM} + \beta_2 T_{lm}^{HIM+FE} + X_{ilm}\gamma + \epsilon_{ilm} \quad (3.1)$$

where Y_{ilm} is the outcome of interest for child i , in preschool l , in triplet m . X_{ilm} is a pre-specified set of control variables added to improve efficiency. 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).

Pre-specified 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 endline in Section 4.2. 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 1000 iterations. 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) which, building on Romano and Wolf (2005) and Romano and Wolf (2010), provides balanced asymptotic control of the family-wise error rate. In running the procedure, we use the 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.

4 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 5, 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 4.1, the measurement models we use and how we estimate the latent factors of interest in the analysis. In Section 4.2, 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; (2) child socio-emotional development. In Section 4.3, 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.

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

4.1 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, for example, Cunha, Heckman, and Schennach (2010); Heckman et al. (2013); Attanasio et al. (2020a); Agostinelli et al. (2021); Heckman, Liu, Lu, and Zhou (2020); Heckman and Zhou (2022)). These techniques combine the information contained in the available measures 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 to 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. Heckman and Zhou, 2022; Heckman et al., 2020; Das and Zajonc, 2010; Singh, 2020). While linear factor models model multiple *continuous* test scores as depending *linearly* on underlying unobserved factors (e.g. Agostinelli et al., 2021; Attanasio et al., 2020a; Cunha et al., 2010; Heckman et al., 2013), IRT models use non-linear 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 (Heckman et al., 2020; Van Der Linden and Hambleton, 1997). 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 either the individual child, teacher, TA or classroom. θ_{id} , however, is not observed directly. Instead, observable item responses, y_{ijd} , 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:

- A1. *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.*

⁹For example, the official scoring algorithm provided with the Woodcock-Munoz 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 that comprised of 1,413 Spanish-speaking children from the USA, six Latin American countries and Spain (Schrack et al., 2005). This norming sample is likely to differ substantially from the children in our sample.

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 5.2, we show evidence that supports assumption (A1) of 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 5.2, we also show our results are not sensitive to relaxing assumptions A2 and A3. Furthermore, Heckman et al. (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 5.2. 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 :

$$Pr(y_{ijd} = 1|\theta_{id}) = g_{jd} + (1 - g_{jd}) \frac{\exp(\alpha_{jd} + \beta_{jd}\theta_{id})}{1 + \exp(\alpha_{jd} + \beta_{jd}\theta_{id})} \quad (4.1)$$

In this set-up, α_{jd} represents an item j ’s difficulty – the higher is 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. g_{jd} is the “pseudo-guessing 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 2-parameter IRT model which is the same as above but restricts the guessing parameter (g_{jd}) to 0.

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

¹¹In 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 multi-step procedure where we first estimate the unrestricted model. We find all items with an estimated guessing parameter greater than 0.05. We then re-estimate 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. Appendix Table C.1 shows which items are estimated to have positive guessing parameters. In Appendix 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 estimated information matrix is not invertible.

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 which models the probability of i having a response of more than k as an ordered logit:

$$Pr(y_{ijd} \geq k | \theta_{id}) = \frac{\exp(\alpha_{jkd} + \beta_{jd}\theta_{id})}{1 + \exp(\alpha_{jkd} + \beta_{jd}\theta_{id})} \quad (4.2)$$

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 (EM) 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 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 re-estimating 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.

4.2 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 socio-emotional development. Child development is a multidimensional construct, as discussed, for instance, in Cunha et al. (2010), Attanasio, Meghir, and Nix (2020b), Attanasio et al. (2020a) and Heckman et al. (2020). Furthermore, preschool has been shown to impact various dimensions of children’s development (Berlinski, Galiani, and Gertler, 2009; Datta Gupta and Simonsen, 2010; Chetty et al., 2011; Heckman et al., 2013; Araujo et al., 2016; Kline and Walters, 2016). Therefore, at both baseline and endline, we used a range of child development assessments that sought to capture children’s development across different domains. The measures we used at endline were richer than those used at baseline. This is because of a combination of cost considerations and the fact that the emphasis of the study is on estimating treatment effects on endline child development, so that baseline measures are primarily useful for checking for balance and increasing the precision of estimated effect sizes.

¹²We do this in order to ensure the associated information matrix is invertible. In Appendix 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%.

4.2.1 Baseline

At baseline, we administered all five subscales of an extended version of the ASQ-3 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 (Jackson-Maldonado et al., 2003, 2013) to measure language development; and the ASQ:SE (Squires, Bricker, and Twombly, 2002) to measure socio-emotional development. These are all parental-report instruments, i.e. 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 4.1. Appendix 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.

4.2.2 Endline

We have endline data for seven child development assessments, each designed to capture a different dimension of child development, including: (1) fluid reasoning; (2) memory for words; (3) expressive language; (4) receptive language; (5) school readiness; (6) inhibitory control; and (7) socioemotional development.¹⁵ Assessments (1) to (3) comprise the relevant scales from the Woodcock-Muñoz-III (WM) tests of cognition and achievement (Schrank et al., 2005), which are Spanish versions of the well-known Woodcock-Johnson tests (Woodcock, 1977). Receptive language was measured using the Spanish version of the Peabody Picture Vocabulary Test (PPVT) – 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), which included only 70 items, chosen through piloting. Inhibitory control, a dimension of executive functioning, was measured using the

¹³The standard ASQ comprises age-specific questionnaires containing six questions for each sub-scale. We extended each age-specific questionnaire by adding the last three non-overlapping items in each sub-scale from the age-specific questionnaire below and the first three non-overlapping items in each sub-scale 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. However, there is strong overlap by age which allows us to use our IRT model to estimate a single factor for each sub-scale. 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).

¹⁴The MacArthur Bates CDI has separate list 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.

¹⁵We also collected measures of sound awareness and concept formation. However, these two tests were too hard for most children so that many 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 (WM cognition 5) and only 5.1% of children progressed past the first nine (out of a total of 18) items on the test of sound awareness (WM achievement 21). Due to this poor performance, we drop these assessments from all analysis

nonverbal Pencil Tapping Task (PTT) (Diamond and Taylor, 1996). Finally, socioemotional development was assessed using the Socio-Emotional Questionnaire in the Ages and Stages Questionnaires (ASQ:SE) (Squires et al., 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 pre-standardized 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 using the publisher’s algorithm based on Item Response Theory (IRT), for the WM tests. 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 endline 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 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 through 6) into a single estimated factor using the procedure outlined in Section 4.1. 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 3.3, we re-estimate 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 mean (median) standard deviation across all bootstrapped samples of a given child’s predicted factor score is 0.15 SD (0.13 SD) for cognitive development. The corresponding figures for socioemotional problems are 0.24 SD and 0.19 SD indicating that these estimates are slightly less precise.

4.3 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 that 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 nappies, brushing teeth and washing hands, 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 (Attanasio et al., 2020a; Heckman and Mosso, 2014), 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 Early Childhood Education qualification), but the TAs are not, a natural split in terms of the allocation of roles between teachers and TAs 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 4.1. 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 (4.2). We repeat this procedure for the learning and development activities. Appendix 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 that 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 social-emotional 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. Due to logistical and budgetary constraints, we only conducted ECERS-R in 172 of the 847 classrooms in our sample.¹⁶

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 the indicators that were due to be answered in all observations and again summarize them using the measurement model described in equation (4.1).¹⁷¹⁸ Appendix Table C.10 presents parameter estimates from this measurement model and suggests almost all items load significantly onto the underlying factor.

5 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 the 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 4. Estimation using these factor scores have the advantage of using information contained in each assessment more efficiently. It also generates impacts which are scaled by the true variance of the underlying factor in the control group, uncontaminated by measurement error induced variability.

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

¹⁶The sub-sample 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 ITERS-R (corresponding assessment for classes of children aged 0–2, 40% of classrooms). At follow-up, we had sufficient budget to collect observations on 211 classrooms in 54 centers. We chose half these classrooms to be the same classrooms we had observed at baseline (randomly chosen) and the other half to be classrooms attended by children in the sample at follow-up (since study children had moved on from their baseline classrooms). This resulted in observations in 172 classrooms with children older than 2 years where we carried out the ECERS-R and 39 classrooms with children aged 0–2 where we 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 cannot be linked to ECERS-R classrooms due to a lack of common items.

¹⁷Each item is formed of around 10 sub-items grouped under the headings “inadequate”, “minimal”, “good” and “excellent” to which the observer must answer “true” or “false”. We followed the official administration procedure, which, unfortunately, turned out to be poorly suited to our context due to stopping rules which resulted in a high number of non-random missing values for items in the “minimal”, “good” and “excellent” categories. We, therefore, only use items from the “inadequate” category in our analysis. While this overcomes the challenge posed by missing data, it implies that the sub-items that make up our quality measures are informative on the absence of poor practices rather than the presence of good ones.

¹⁸To increase the sample size for estimating the measurement system parameters, we pool ECERS-R measures from baseline and endline giving a total sample of 296 observations.

component to the HIM program (i.e. the difference between the HIM and HIM+FE programs). Appendix Table B.6 shows estimated coefficients on the control variables.

5.1 Child Development

Columns 1–6 of Table 2 show estimates of the impacts of the interventions on children’s performance in each of the child cognitive 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.

Table 2: Impacts on Child Development Assessments

	(1) Fluid Reasoning	(2) Memory for words	(3) Expressive Language	(4) School Readiness	(5) Receptive Language	(6) Inhibitory Control	(7) Socioemotional Problems
HIM	-0.083 (0.462) <i>p</i> =0.877 <i>q</i> =0.971	1.474 (2.992) <i>p</i> =0.632 <i>q</i> =0.961	-0.749 (1.366) <i>p</i> =0.575 <i>q</i> =0.961	0.283 (0.833) <i>p</i> =0.724 <i>q</i> =0.971	-1.447 (1.510) <i>p</i> =0.330 <i>q</i> =0.854	0.253 (0.367) <i>p</i> =0.502 <i>q</i> =0.961	0.723 (2.687) <i>p</i> =0.800 <i>q</i> =0.971
HIM+FE	0.918* (0.487) <i>p</i> =0.057 <i>q</i> =0.261	4.621 (2.952) <i>p</i> =0.115 <i>q</i> =0.372	2.407* (1.261) <i>p</i> =0.057 <i>q</i> =0.261	1.883*** (0.671) <i>p</i> =0.003 <i>q</i> =0.028	0.737 (1.241) <i>p</i> =0.550 <i>q</i> =0.759	0.360 (0.412) <i>p</i> =0.384 <i>q</i> =0.759	-1.920 (2.754) <i>p</i> =0.476 <i>q</i> =0.759
Difference	1.000** (0.410) <i>p</i> =0.014 <i>q</i> =0.073	3.147 (2.448) <i>p</i> =0.185 <i>q</i> =0.474	3.156*** (1.104) <i>p</i> =0.004 <i>q</i> =0.026	1.601** (0.717) <i>p</i> =0.030 <i>q</i> =0.113	2.184* (1.216) <i>p</i> =0.076 <i>q</i> =0.249	0.107 (0.336) <i>p</i> =0.771 <i>q</i> =0.771	-2.643 (2.345) <i>p</i> =0.257 <i>q</i> =0.474
N	1073	1073	1073	1074	1074	1074	1074
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 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 et al. (2019). Clustered standard errors (bootstrapped) 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 Section 4.2.

* *p* < 0.1, ** *p* < 0.05, *** *p* < 0.01

We do, however, find evidence that the teacher professional development training program combined with the government improvement program (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 (“Difference” row), we see statistically significant

improvements across four measures (the three above and 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, combined, the HIM+FE program led to improvements of 0.16 of a standard deviation (SD) relative to the pure control, with a p -value of 0.025. The final 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 ($p = 0.005$).

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 under-powered 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 Section 4.2).

Table 3: Impacts on Cognitive and Socioemotional Factor Scores

	(1) Cognitive Development	(2) Socioemotional Problems
HIM	-0.008 (0.087) $p=0.925$	-0.014 (0.155) $p=0.933$
HIM+FE	0.161** (0.073) $p=0.025$	-0.155 (0.174) $p=0.382$
Difference	0.169*** (0.060) $p=0.005$	-0.141 (0.148) $p=0.341$
N	1074	1074

Note. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. We re-estimate 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 scaled so that the underlying latent factor has a mean of 0 and standard deviation of 1 in the control group. All factors constructed as described in Section 4.2.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.2 Robustness

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

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 are dependent on specific modelling 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 the underlying latent factors are normally distributed in the control group. We use both non-parametric (Columns 1 and 8) and semi-parametric methods (Columns 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 (Columns 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 (Column 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 (Column 6).¹⁹ 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 (Column 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 et al. (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.

Controls: In Appendix Table B.8, we show that our results are also not sensitive to the control variables we include. Even controlling 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, 0.073 and 0.016. Controlling for age and gender yields p -values of 0.048 and 0.017, 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).

¹⁹The exploratory factor analysis yields a first factor with an eigenvalue of 2.99, while the second factor has an eigenvalue of just 0.28.

Sample: All of our main conclusions hold when we also include younger children (below 48 months at baseline) 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 (Appendix Table B.9). Furthermore, in Appendix 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 comparison 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 (Appendix Table B.11).

5.3 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.²⁰ Estimates in column 1 of Table 4 show first 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 column 3).

5.4 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 measured by a factor aggregating the MacArthur-Bates CDI and the ASQ-3 administered at baseline (see Section 4), as having “higher baseline development”.²¹ We find that while HIM+FE resulted in a large and statistically 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 column 2 of Table 4). The impacts on socioemotional development are not different from zero for either group (column 4).

²⁰This wealth index was constructed by summarizing information about whether the household owned at least one of 8 different assets (including, for example, a car, a TV, a washing machine) through factor analysis.

²¹When controlling for baseline child development in the main analysis, we included each of the sub-scales of ASQ-3 and MacArthur-Bates CDI separately. However, for this heterogeneity analysis, we want to summarize all of these sub-scales 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.

Table 4: Heterogeneity by Wealth and Baseline Child Development

	<i>Cognitive Development</i>		<i>Socio-Emotional Problems</i>	
	(1)	(2)	(3)	(4)
HIM	0.028 (0.112) <i>p</i> =0.814	0.103 (0.117) <i>p</i> =0.374	0.007 (0.210) <i>p</i> =0.971	0.095 (0.204) <i>p</i> =0.645
HIM+FE	0.286*** (0.090) <i>p</i> =0.001	0.301*** (0.092) <i>p</i> =0.002	-0.115 (0.203) <i>p</i> =0.559	-0.146 (0.210) <i>p</i> =0.486
HIM x Wealthier	-0.070 (0.139) <i>p</i> =0.615		-0.046 (0.199) <i>p</i> =0.825	
HIM+FE x Wealthier	-0.252** (0.125) <i>p</i> =0.045		-0.088 (0.241) <i>p</i> =0.716	
HIM x Higher BL dev		-0.227* (0.131) <i>p</i> =0.081		-0.208 (0.284) <i>p</i> =0.472
HIM+FE x Higher BL dev		-0.291** (0.113) <i>p</i> =0.013		-0.015 (0.266) <i>p</i> =0.964
Difference	0.258*** (0.076) <i>p</i> =0.001	0.197** (0.087) <i>p</i> =0.018	-0.122 (0.162) <i>p</i> =0.487	-0.241 (0.163) <i>p</i> =0.136
Difference x Wealthier	-0.182* (0.098) <i>p</i> =0.063		-0.042 (0.241) <i>p</i> =0.873	
Difference x Higher BL dev		-0.063 (0.114) <i>p</i> =0.578		0.193 (0.266) <i>p</i> =0.475
N	1074	1074	1074	1074

Note. Two-sided *p*-values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. We re-estimate 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 (columns 1 and 3) and baseline child development (columns 2 and 4). All factors scaled so the underlying latent factor has a mean of 0 and standard deviation of 1 in the control group. All factors constructed as described in Section 4. “Wealthier” implies child’s household had above-median value of household asset index at baseline. “Higher BL dev” implies child had above-median baseline child development as measured by the factor score discussed in Section 5.4.

* *p* < 0.1, ** *p* < 0.05, *** *p* < 0.01

6 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 have 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 work overtime have 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 instruct any TAs they have to allocate TA classroom time. Teachers in HIs tend to have a high degree of autonomy over how they manage their class and so it is reasonable to assume that they choose how their time and that of their TAs is 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 tend 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 as a proxy for total time that they spend on their job, data from the TSEEQ to capture allocation of time between learning and caring activities, and ECERS-R data to measure quality of teaching by teachers within the classroom (see Section 4 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 pre-school level (e.g. average overtime of all teachers in the pre-school). 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 (column 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 (columns 5 through 7) although the precision decreases in some cases. In the last column, we estimate the same regression as in column 7 but restricting the sample to children in the control 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 which we formalize in an economic model presented in Appendix E. As motivated above, we consider teachers' choices about how much time

Table 5: Correlations between Child Cognitive Development and Teacher-Reported Activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Learning Activities	0.084** (0.033) $p=0.012$				0.097* (0.050) $p=0.054$	0.090* (0.047) $p=0.056$	0.078 (0.056) $p=0.167$	0.100 (0.101) $p=0.329$
Caring Activities		0.021 (0.030) $p=0.482$			-0.015 (0.038) $p=0.703$	-0.017 (0.038) $p=0.655$	-0.017 (0.041) $p=0.689$	-0.035 (0.082) $p=0.671$
Overtime			0.048* (0.027) $p=0.081$			0.041 (0.027) $p=0.126$	0.021 (0.033) $p=0.524$	0.051 (0.098) $p=0.608$
ECERS Quality				0.209* (0.116) $p=0.076$			0.196* (0.112) $p=0.083$	0.193 (0.190) $p=0.318$
N	1074	1074	1074	726	1074	1074	726	249

Note. Two-sided p -values and standard errors are clustered at the HI level. Table presents OLS 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 Section 4. 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 only.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

to spend in the classroom, as well as how to allocate this time between care and learning activities. In making these choices, we assume that teachers care about both child development and leisure and that they make these choices 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 towards 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 their 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 to be 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 amount of time teachers spend on classroom activities will only depend on 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 (Column 1), frequency with which learning and care activities are undertaken by the teachers (Columns 2 and 3) and the TAs (Columns 4 and 5), and, finally, quality of teaching delivered by the teachers (Column 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. 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 quality of teaching delivered by the teachers; HIM has no impact on psychologists’ assessment of overall quality of teaching by the teachers during classroom observation.²² 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.

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 place 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 behaviour. 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.²³ 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

²²Note that we only have these measures for 172 out of the 841 classrooms in the sample (see Section 4.3 for details).

²³Our 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).

Table 6: Impacts on Teachers’ and TAs’ Behavior

	<i>Teachers’ time</i> Overtime	<i>Teachers’ activities</i> Learning	Care	<i>TA’s activities</i> Learning	Care	<i>Observed Teaching</i> <i>Quality</i>
	(1)	(2)	(3)	(4)	(5)	(6)
HIM only	-0.450** (0.218) $p=0.035$	-0.359** (0.171) $p=0.038$	-1.020*** (0.270) $p=0.002$			0.008 (0.095) $p=0.936$
FE+HIM	-0.033 (0.275) $p=0.911$	-0.045 (0.162) $p=0.772$	-0.758** (0.270) $p=0.012$	0.186 (0.217) $p=0.398$	0.019 (0.248) $p=0.923$	0.180** (0.093) $p=0.050$
Difference	0.417 (0.270) $p=0.129$	0.314** (0.152) $p=0.040$	0.262 (0.241) $p=0.268$			0.172* (0.105) $p=0.089$
N	841	841	839	254	254	172

Note. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. We re-estimate 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 column (6) where this is not possible because since we do not have ECERS measures for all HIs, there is one city that doesn’t 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 standard deviation of 1 in the control group (HIM group in the case of TA activities). All factors constructed as described in Section 4.3.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

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, not mutually exclusive, 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 their time towards these activities.²⁴ 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 ($p = 0.050$), with the difference between the HIM and HIM+FE arms being similar in magnitude and statistical significance.²⁵

²⁴In Appendix 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 they would exploit their comparative advantage in caring activities. Thus, the fact that we find no significant impacts on TA time use (Columns 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 don’t draw strong conclusions.

²⁵If we were to go further and assume that the quality of teaching observed during the ECERS assessment summarized all aspects of classroom quality relevant to child development and was the only channel through which the treatments impacts 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 Appendix 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

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.

7 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 the striking finding that a national, costly, 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 17% of a standard deviation of the control group and especially large benefits of 29% of a standard deviation for the more disadvantaged children in the sample.

These are non-negligible impacts. To the extent that credible comparisons can be made between studies, 17% of a standard deviation 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 USA, 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 which 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 zero impact of hiring 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

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, since teachers were always present in the classroom while the observation was happening, ECERS 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.

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/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 et al. (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 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 these. 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.

References

- Agostinelli, Francesco, Ciro Avitabile, and Matteo Bobba. 2021. “Enhancing Human Capital at Scale.” Discussion Paper 14192, IZA.
- Agostinelli, Francesco and Matthew Wiswall. 2016. “Estimating the Technology of Children’s Skill Formation.” Working Paper 22442, National Bureau of Economic Research.
- 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.” *The Quarterly Journal of Economics* 131 (3):1415–1453.
- Araujo, Maria 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, chap. 4. New York: Palgrave Macmillan US, 91–121.

- Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020a. “Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia.” *American Economic Review* 110 (1):48–85.
- Attanasio, Orazio, Costas Meghir, and Emily Nix. 2020b. “Human Capital Development and Parental Investment in India.” *The Review of Economic Studies* 87:2511–2541.
- Banerjee, A, R Banerji, J Berry, E Duflo, H Kannan, S Mukerji, M Shotland, and Walton W. 2017. “From proof of concept to scalable policies: challenges and solutions, with an application.” *Journal of Economic Perspectives* 31 (4):73–102.
- Banerjee, A, S Cole, E Duflo, and L Linden. 2007. “Remedying education: Evidence from two randomized experiments in India.” *The Quarterly Journal of Economics* .
- Bartlett, Maurice S. 1937. “The statistical conception of mental factors.” *British journal of Psychology* 28 (1):97.
- Behrman, Jere R, John Hoddinott, John A Maluccio, Erica Soler-Hampejsek, Emily L Behrman, Reynaldo Martorell, Manuel Ramírez-Zea, and Aryeh D Stein. 2014. “What determines adult cognitive skills? Influences of pre-school, school, and post-school experiences in Guatemala.” *Latin American Economic Review* 23 (1).
- Berlinski, Samuel, Sebastian Galiani, and Paul Gertler. 2009. “The effect of pre-primary education on primary school performance.” *Journal of Public Economics* 93 (1-2):219–234.
- Berlinski, Samuel, Sebastian Galiani, and Marco Manacorda. 2008. “Giving children a better start: Preschool attendance and school-age profiles.” *Journal of Public Economics* 92 (5-6):1416–1440.
- 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 Research Quarterly* 47:418–431.
- 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.” *Social Science & Medicine* 97 (C):241–249.
- 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, A Lord. 1968. “Some latent trait models and their use in inferring an examinee’s ability.” *Statistical theories of mental test scores* .
- Bock, R Darrell and Murray Aitkin. 1981. “Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm.” *Psychometrika* 46 (4):443–459.

- Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng, and Justin Sandefur. 2018. “Experimental evidence on scaling up education reforms in Kenya.” *Journal of Public Economics* 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 and commentaries.” *Social Policy Report* 25:1–31.
- 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):339–357.
- Caucutt, Elizabeth M, Lance Lochner, and Youngmin Park. 2017. “Correlation, Consumption, Confusion, or Constraints: Why do Poor Children Perform so Poorly?” *Scandinavian Journal of Economics* 119 (1):102–147.
- 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.” *Quarterly Journal of Economics* 126 (4):1593–1660.
- Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg. 2018. “Who Benefits from Universal Child Care? Estimating Marginal Returns to Early Child Care Attendance.” *Journal of Political Economy* 126 (6):2356–2409.
- Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. “Estimating the Technology of Cognitive and Noncognitive Skill Formation.” *Econometrica* 78 (3):883–931.
- Cunha, Flávio, Irma Elo, and Jennifer Culhane. 2022. “Maternal subjective expectations about the technology of skill formation predict investments in children one year later.” *Journal of Econometrics* 231:3–32.
- Danzer, Virginia A, Mary Frances Gerber, Theresa M Lyons, and Judith K Voress. 1991. *Daberon 2: Screening for School Readiness*. Pro-Ed (Firm).
- Das, Jishnu and Tristan Zajonc. 2010. “India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement.” *Journal of Development Economics* 92 (2):175–187.
- Datta Gupta, Nabanita and Marianne Simonsen. 2010. “Non-cognitive child outcomes and universal high quality child care.” *Journal of Public Economics* 94 (1-2):30–43.
- de Chaisemartin, Clement and Jaime Ramirez-Cuellar. 2020. “At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?” Working Paper 27609, National Bureau of Economic Research.
- Deaton, Angus. 2010. “Instruments, randomization, and learning about development.” *Journal of Economic Literature* 48 (2):424–455.

- Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1 (3):111–134.
- Diamond, A and C Taylor. 1996. "Development of an aspect of executive control: development of the abilities to remember what I said and to "do as I say, not as I do"." *Developmental psychobiology* 29 (4):315–334.
- Duflo, Annie, Jessica Kiessel, and Adrienne M Lucas. 2020. "External Validity: Four Models of Improving Student Achievement." Working Paper 27298, National Bureau of Economic Research.
- Dunn, Lloyd M, Eligio R Padilla, Delia E Lugo, and Leota M Dunn. 1986. *Test de Vocabulario en Imágenes Peabody (TVIP)*. AGS Circle Pines, MN.
- Elango, Sneha, Jorge Luis Garcia, James Heckman, and Andrés Hojman. 2015. "Early Childhood Education." In *Economics of Means-Tested Transfer Programs in the United States, Volume 2*. National Bureau of Economic Research, Inc, 235–297.
- Engle, Patrice L, Lia C H Fernald, Harold Alderman, Jere Behrman, Chloe O’Gara, Aisha Yousafzai, Meena Cabral de Mello, Melissa Hidrobo, Nurper Ulkuer, Ilgi Ertem, and Selim Iltus. 2011. "Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries." *The Lancet* 378 (9799):1339–1353.
- Evans, David K and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *The World Bank Research Observer* 31 (2):242–270.
- Felfe, Christina and Rafael Lalive. 2018. "Does early child care affect children’s development?" *Journal of Public Economics* 159 (January):33–53.
- Garces, Eliana, Duncan Thomas, and Janet Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92 (4):999–1012.
- Gerber, Susan B, Jeremy D Finn, Charles M Achilles, and Jayne Boyd-Zaharias. 2001. "Teacher Aides and Students’ Academic Achievement." *Educational Evaluation and Policy Analysis* 23 (2):123–143.
- Glewwe, P. and K. Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." *Handbook of the Economics of Education* 5:653–743.
- 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 17554, National Bureau of Economic Research.
- Hallam, R, B Rous, S Riley-Ayers, and D Epstein. 2011. *Teacher survey of early education quality*. New Brunswick, NJ: NIEER.

- Hanushek, Eric A. 1999. "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects." *Educational Evaluation and Policy Analysis* 21 (2):143–163.
- Harms, T, R M Clifford, and D Cryer. 1998. *Early Childhood Environment Scale-Revised Edition*. New York: Teachers College Press.
- Havnes, Tarjei and Magne Mogstad. 2015. "Is universal child care leveling the playing field?" *Journal of Public Economics* 127:100–114.
- Heckman, J and J Zhou. 2021. "Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning." .
- Heckman, James, Bei Liu, M. Lu, and Jin Zhou. 2020. "The Impacts of a Prototypical Home Visiting Program on Child Skills." Working Paper 27356, National Bureau of Economic Research.
- Heckman, James, Rodrigo Pinto, and Peter Savelyev. 2013. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes." *American Economic Review* 103 (6):2052–2086.
- Heckman, James and Jin Zhou. 2022. "Measuring Knowledge and Learning." Working Paper 29990, National Bureau of Economic Research, Cambridge, MA.
- Heckman, James J. 1992. "Randomization and Social Policy Evaluation." *Evaluating welfare and training programs* :201.
- 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." *Journal of Public Economics* 94 (1-2):114–128.
- Heckman, James J and Stefano Mosso. 2014. "The Economics of Human Development and Social Mobility." *Annual Review of Economics* 6 (1):689–733.
- Ichino, Andrea, Margherita Fort, and Giulio Zanella. 2019. "Cognitive and Non-Cognitive Costs of Daycare 0-2 for Children in Advantaged Families." *Journal of Political Economy* :704075.
- Jackson-Maldonado, Donna, Virginia A Marchman, and Lia CH Fernald. 2013. "Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories." *Applied Psycholinguistics* 34 (04):837–868.
- Jackson-Maldonado, Donna, Donna J. Thal, Larry Fenson, Virginia A. Marchman, Tyler Newton, Barbara T. Conboy, and Larry Fenson. 2003. *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual*. Baltimore: Brookes Publishing Co.
- 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 Education and Development* 31 (1):1–26.

- Kaiser, Henry F. 1960. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement* 20 (1):141–151.
- Kline, Patrick and Christopher R Walters. 2016. "Evaluating Public Programs with Close Substitutes: The Case of Head Start." *The Quarterly Journal of Economics* 131 (4):1795–1848.
- Krueger, Alan B. 1999. "Experimental estimates of education production functions." *Quarterly Journal of Economics* 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." *Economic Journal* 111 (468):1–28.
- List, John A., Azeem M. Shaikh, and Yang Xu. 2019. "Multiple hypothesis testing in experimental economics." *Experimental Economics* 22:773–793.
- Neuman, Michelle J. and Lynette Okeng'o. 2019. "Early childhood policies in low- and middle-income countries." *Early Years* 39:223–228.
- Noboa-Hidalgo, Grace E and Sergio S Urzúa. 2012. "The Effects of Participation in Public Child Care Centers: Evidence from Chile." *Journal of 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." *Economics of Education Review* 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." *Journal of Development Economics* 133 (August 2017):448–467.
- Peisner-Feinberg, Ellen S, Margaret R Burchinal, Richard M Clifford, Mary L Culkin, Carollee Howes, Sharon Lynn Kagan, and Noreen Yazejian. 2001. "The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade." *Child development* 72 (5):1534–1553.
- Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning*, vol. 123. Brookings Institution Press.
- Romano, Joseph and Michael Wolf. 2010. "Balanced control of generalized error rates." *Annals of Statistics* 38 (1):598–633.
- Romano, Joseph P and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73 (4):1237–1282.
- Rosero, Jose José and Hessel Oosterbeek. 2011. "Trade-offs between Different Early Childhood Interventions: Evidence from Ecuador." Discussion Paper 11-102/3, Tinbergen Institute.

- 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 Science* (January):1–15.
- Sato, Ryuzo and Tetsunori Koizumi. 1973. "On the Elasticities of Substitution and Complementarity." *Oxford Economic Papers* 25 (1):44–56.
- 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 Publishing.
- Singh, Abhijeet. 2020. "Learning More with Every Year: School Year Productivity and International Learning Divergence." *Journal of the European Economic Association* 18 (4):1770–1813.
- Squires, Jane, D Bricker, and E Twombly. 2002. *Ages & Stages Questionnaires: Social-emotional*. Brookes Pub. Co.
- . 2009. *Ages & Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed, Child-Monitoring System*. Baltimore: Paul H. Brookes Publishing Co.
- Sylva, Kathy, Iram Siraj-Blatchford, Brenda Taggart, Pam Sammons, Edward Melhuish, Karen Elliot, and Vasiliki Totsika. 2006. "Capturing quality in early childhood through environmental rating scales." *Early Childhood Research Quarterly* 21 (1):76–92.
- Van Der Linden, Wim J and Ronald K Hambleton. 1997. *Handbook of Modern Item Response Theory*. New York, NY: Springer New York.
- Wolf, Sharon. 2018. "Impacts of Pre-Service Training and Coaching on Kindergarten Quality and Student Learning Outcomes in Ghana." *Studies in Educational Evaluation* 59:112–123.
- Woodcock, Richard W. 1977. "Woodcock-Johnson Psycho-Educational Battery. Technical Report." .
- World Bank. 2018. "Learning to Realize Education's Promise." *World Development Report 2018* .
- Yoshikawa, Hirokazu, Diana Leyva, Catherine E Snow, Ernesto Treviño, M Clara Barata, Christina Weiland, Celia J Gomez, Lorenzo Moreno, Andrea Rolla, Nikhit D'Sa, and Mary Catherine Arbour. 2015. "Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes." *Developmental Psychology* 51 (3):309–322.
- Yoshikawa, 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." *Social Policy Report* 31 (1):1–36.

A Costs of the two programs

A.1 Costs of HIM, the government improvement program

The HIM program comprised increasing the the amount that HIs received per student per year from roughly \$1000 to \$1300, a substantial 30% increase in per-child investment.

A.2 Costs of Pedagogical Training Program

Here we provide more details on the costs associated with the FE teacher professional development training program, which we argue, was the key ingredient for generating the child development impacts that we find. FE provided us with the total cost of this component (COP 419,546,284, or USD 233,081 at February 2013 exchange rate of 1,800 COP/USD). With this budget, FE provided completed training for 99 teachers from the 40 HIs in the HIM+FE treatment arm.²⁶ This represented 32% of teachers who worked in those 40 HIs. We thus estimate that the initial one-off cost to roll out the professional development program to new HIs, at the same intensity as that leading to the impacts we see in this study (i.e. training 32% of teachers), would be USD 5,827 per HI or USD 35 per child attending an HI.

However, it is unreasonable to assume that the same intensity of training would be required year after year for FE to sustain its impacts on successive cohorts. Rather, we calculate the costs of maintaining a ratio of training 32% of staff, which implies providing training for 32% of new staff, and the costs of providing a yearly refresher training to all teachers who have already been trained which we assume would cost 25% of the costs of the full training. Given these assumptions, we estimate that the ongoing cost per center of maintaining the results of the professional development training program would be USD 2206 per year and the cost per child would be USD 13 per year. All data, assumptions and formulae used in these calculations are shown in Table A.1.

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

²⁶More teachers began training however, for consistency, we calculate costs relative to the number completing. Presuming the drop-out rates seen during the study are similar to what they would be if the program were scaled, this makes no difference. We also note that in some cases other staff (headteachers, teaching assistants etc) also completed the training. To be as conservative as possible in calculating costs we simply calculate cost per teacher completing rather than cost per person. This means that our projected costs also allow the same proportion of other staff to receive training as they did in the trial.

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

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

B Additional Tables

Table B.1: Attrition by Treatment Status

	Extended Sample	Analysis Sample
	All kids	Kids 48 months +
HIM	0.000 (0.020) $p=0.991$	-0.014 (0.017) $p=0.430$
FE+HIM	0.012 (0.018) $p=0.487$	-0.013 (0.018) $p=0.477$
Difference	0.012 (0.016) $p=0.465$	0.001 (0.015) $p=0.958$
N	1987	1149
Control mean	0.921	0.941

Notes: Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. Estimates from a regression of a dummy indicating non-attrition on treatment status. Column (1) assesses attrition amongst all children with baseline data and regresses an indicator that we have at least one endline child development assessment on treatment status. Column (2) focuses on children above 48 months of age at the time when endline assessments were carried out in their HI. We regress an indicator of whether these assessments were collected on treatment status for all children whom would have been 48 months or older on the median date of assessments (assessments only lasted 2-3 days on average per HI) and date of birth.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: Compliance with HIM hiring recommendations

	(1) # TAs per 50 kids	(2) # FTE TAs per 50 kids	(3) # SEs per 200 kids	(4) # FTE SEs per 200 kids	(5) # NEs per 200 kids	(6) # FTE NEs per 200 kids	(7) # TAs per teacher
HIM	0.863*** (0.050) $p<0.001$	0.930*** (0.069) $p<0.001$	0.946*** (0.135) $p<0.001$	0.722*** (0.093) $p<0.001$	1.114*** (0.127) $p<0.001$	0.458*** (0.068) $p<0.001$	0.441*** (0.033) $p<0.001$
FE+HIM	0.871*** (0.043) $p<0.001$	0.952*** (0.054) $p<0.001$	0.767*** (0.152) $p<0.001$	0.662*** (0.083) $p<0.001$	1.077*** (0.126) $p<0.001$	0.495*** (0.085) $p<0.001$	0.410*** (0.022) $p<0.001$
Difference	0.008 (0.056) $p=0.898$	0.022 (0.084) $p=0.798$	-0.179 (0.129) $p=0.173$	-0.060 (0.075) $p=0.438$	-0.037 (0.128) $p=0.770$	0.037 (0.069) $p=0.604$	-0.031 (0.036) $p=0.404$
N	120	120	120	120	120	120	120
Control mean	0.073	0.084	0.552	0.291	0.319	0.110	0.035

Notes: Table shows impacts on the number of TAs, Socioemotional Experts (SEs) and Nutritional Experts (NEs) present in the HI at endline. FTE refers to full-time equivalent to take into account part time working and overtime. Column (7) is the TA to teacher ratio. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. No control variables are used.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.3: Child Development Assessments

Dimension	Instrument used	Acronym
Fluid reasoning	Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 12	WM12
Memory for words	Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 17	WM17
Expressive language	Woodcock-Muñoz: Pruebas de Aprovechamiento – 14	WM14
Receptive Language	Test de Vocabulario en Imágenes de Peabody	TVIP
School readiness	Daberon-II Screening for School Readiness	DAB
Inhibitory control	Pencil Tapping Task	PTT
Socio-Emotional Development	ASQ:SE: Interaction with People	ASQ:SE
Concept formation*	Woodcock-Muñoz: Pruebas de Habilidades Cognitivas – 5	WM5
Sound Awareness*	Woodcock-Muñoz: Pruebas de Aprovechamiento – 21	WM21

Notes: *: These two assessments performed very poorly and were dropped from all analysis. They were too hard for most children so that most did not progress past the initial few items, leaving very little information. Specifically, only 25.9% of children progressed past the first five items in the test of concept formation (WM5) and only 5.1% of children progressed past the first nine items on the test of sound awareness (WM21).

Table B.4: Correlation of Child Development Assessments with Age Baseline Child Development, and Socio-Economic Status

	Baseline Child Development										
	Age	Problem Solving	Communication	Gross Motor	Fine Motor	Socio-Individual	MacArthur-Bates	Socioemotional	Wealth index	Mother's education	N
Fluid Reasoning	0.170***	0.224***	0.079***	0.075**	0.156***	0.114***	0.243***	-0.023	0.162***	0.192***	1,073
Memory for Words	0.202***	0.231***	0.065**	0.116***	0.137***	0.075**	0.207***	-0.057*	0.121***	0.095***	1,073
Expressive Language	0.166***	0.238***	0.102***	0.097***	0.192***	0.058*	0.232***	-0.075**	0.302***	0.273***	1,073
School Readiness	0.290***	0.270***	0.101***	0.138***	0.201***	0.138***	0.285***	-0.075**	0.225***	0.209***	1,074
Receptive Language	0.224***	0.209***	0.098***	0.111***	0.171***	0.099***	0.242***	-0.093***	0.260***	0.248***	1,074
Inhibitory Control	0.188***	0.154***	0.04	0.084**	0.093***	0.075**	0.149***	-0.032	0.113***	0.121***	1,074
Socio-Emotional Problems	-0.038	-0.150***	-0.065**	-0.098***	-0.106***	-0.137***	-0.146***	0.112***	-0.217***	-0.139***	1,074

Notes: Correlation coefficient significantly different from zero at: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Wealth index constructed using factor analysis of baseline asset ownership. Years of education of mother as measured at baseline. All child development scores are scored using algorithms recommended by the test publishers as described in Section 4.2. In all cases, these scores are not standardized for age.

Table B.5: Contemporaneous Correlation of Child Development Assessments

	Fluid Reasoning	Memory for Words	Expressive Language	School Readiness	Receptive Language	Inhibitory Control	Socio-Emotional Problems
Fluid Reasoning	1						
Memory for Words	0.342***	1					
Expressive Language	0.512***	0.324***	1				
School Readiness	0.551***	0.476***	0.624***	1			
Receptive Language	0.468***	0.341***	0.645***	0.636***	1		
Inhibitory Control	0.271***	0.320***	0.290***	0.479***	0.334***	1	
Socio-Emotional Problems	-0.156***	-0.0878***	-0.216***	-0.221***	-0.225***	-0.148***	1

Notes: Correlation coefficient significantly different from zero at: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. All child development scores are scored using algorithms recommended by the test publishers as described in Section 4.2. In all cases, these scores are not standardized for age.

Table B.6: Coefficient Estimates on Control Variables

	(1) Cognitive Development	(2) Socioemotional Problems
HIM	-0.008 (0.087) <i>p</i> =0.925	-0.014 (0.155) <i>p</i> =0.933
HIM+FE	0.161** (0.073) <i>p</i> =0.025	-0.155 (0.174) <i>p</i> =0.382
BL MacArthur Bates (older kids)	0.205*** (0.039) <i>p</i> =0.000	-0.043 (0.056) <i>p</i> =0.455
BL MacArthur Bates (younger kids)	0.161*** (0.058) <i>p</i> =0.004	-0.112 (0.086) <i>p</i> =0.188
=1 if older	0.070 (0.114) <i>p</i> =0.523	0.107 (0.220) <i>p</i> =0.611
BL ASQ Communication	0.206*** (0.042) <i>p</i> =0.000	-0.220*** (0.059) <i>p</i> =0.000
BL ASQ Gross Motor	-0.036 (0.040) <i>p</i> =0.362	0.001 (0.054) <i>p</i> =0.982
BL ASQ Fine Motor	0.018 (0.036) <i>p</i> =0.600	0.028 (0.062) <i>p</i> =0.654
BL ASQ Problem Solving	0.082** (0.036) <i>p</i> =0.022	0.020 (0.063) <i>p</i> =0.752
BL ASQ Socio-Individual	-0.032 (0.050) <i>p</i> =0.517	-0.097 (0.073) <i>p</i> =0.184
BL ASQ Socioemotional	-0.003 (0.041) <i>p</i> =0.944	0.240*** (0.075) <i>p</i> =0.001
Male	0.015 (0.049) <i>p</i> =0.736	0.053 (0.082) <i>p</i> =0.531
Age in Months	-0.799 (0.857) <i>p</i> =0.330	2.710 (1.724) <i>p</i> =0.108
Age in Months squared	0.009 (0.008) <i>p</i> =0.272	-0.026 (0.016) <i>p</i> =0.106
N	1074	1074
City Dummies	X	X
Constant	X	X

Note. Table shows coefficient estimates on the control variables for the specification used to estimate main impacts on on the cognitive and socioemotional factors (Table 3). Since younger (30 months and younger at BL) and older child did different versions of the MacArthur Bates at BL, we control separately for each, replacing missings with the mean value for each and then include an indicator for whether or not the child took the “older” version of the assessment. Two-sided *p*-values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). We re-estimate the measurement system on each bootstrapped sample. Clustered standard errors (bootstrapped) in parentheses.

* *p* < 0.1, ** *p* < 0.05, *** *p* < 0.01

Table B.7: Sensitivity of Estimates to Choices Made in Estimation of Factor Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	<i>Cognitive development</i>				<i>Socioemotional problems</i>				
HIM	-0.006 (0.087) <i>p</i> =0.931	-0.006 (0.087) <i>p</i> =0.942	-0.007 (0.085) <i>p</i> =0.917	0.003 (0.087) <i>p</i> =0.967	-0.023 (0.097) <i>p</i> =0.806	-0.022 (0.083) <i>p</i> =0.784	-0.013 (0.094) <i>p</i> =0.880	-0.015 (0.157) <i>p</i> =0.923	-0.014 (0.155) <i>p</i> =0.922
HIM+FE	0.165** (0.075) <i>p</i> =0.019	0.164** (0.075) <i>p</i> =0.019	0.159** (0.073) <i>p</i> =0.023	0.171** (0.076) <i>p</i> =0.018	0.171** (0.077) <i>p</i> =0.020	0.149** (0.067) <i>p</i> =0.020	0.187** (0.078) <i>p</i> =0.012	-0.158 (0.177) <i>p</i> =0.375	-0.154 (0.175) <i>p</i> =0.382
Difference	0.171*** (0.066) <i>p</i> =0.007	0.170*** (0.066) <i>p</i> =0.007	0.166*** (0.066) <i>p</i> =0.008	0.168*** (0.066) <i>p</i> =0.007	0.194*** (0.072) <i>p</i> =0.007	0.170*** (0.062) <i>p</i> =0.007	0.200*** (0.076) <i>p</i> =0.008	-0.144 (0.152) <i>p</i> =0.353	-0.140 (0.149) <i>p</i> =0.353
N	1074	1074	1074	1074	1073	1073	1071	1074	1074
Model type:									
IRT	Y	Y	Y	Y				Y	Y
IRT + linear (Confirm.)					Y				
IRT + linear (Explor.)						Y			
Linear (Confirm.)							Y		
Density of factors	EH	Davidian	Normal	Normal	Normal	Normal	NA	EH	Davidian
Error dependence	iid	iid	iid	iid	Nested	Nested	iid	iid	iid
Inc. items with little variation	N	N	Y	N	N	N	N	NA	NA
Inc. guessing parameters for all binary items	N	N	N	Y	N	N	N	NA	NA

Note. Table shows impacts on cognitive and socioemotional factors estimated relaxing different assumptions:

- Columns (1), (2), (8) and (9) don't impose normality on the underlying distribution, but instead use an empirical histogram approach described by Bock and Aitkin (1981) (columns 1 and 8) and the Davidian semi-parametric approach (Columns 2 and 9).
- Column (3) includes binary items with little variation (those where more than 90% of responses take a single value).
- Column (4) includes guessing parameters for every binary item.
- Column (5) implements a nested measurement error structure. In particular, we allow for some correlation between measurement errors between items (j) that come from the same measure (m). We still have a dedicated measurement system, so all cognitive items depend only on the one underlying cognitive factor. For instance, for binary items with no guessing parameter we have $y_i^{mj} = \mathbf{1}(\alpha^{mj} + \beta^{mj}\theta_i + \varepsilon_i^{mj} > 0)$, $\forall j, m$. Unlike in our main analysis, we do not want to assume that ε_i^{mj} is iid across all i and j . Instead, we want to allow for some correlation between ε_i^{mj} within m . We assume that this correlation takes the following form: $\varepsilon_i^{mj} = \lambda^m \beta^{mj} v_i^m + \omega_i^{mj}$ where ω_i^{mj} is assumed to be iid logistic (with mean 0 and variance $\frac{\pi^2}{3}$) across i and j within each m and independent from v_i^m . v_i^m is iid across i and m with mean 0 and variance 1. λ^m captures the degree of correlation of measurement errors within each measure m and can vary across m . Substituting in, we have: $y_i^{mj} = \mathbf{1}(\alpha^{mj} + \beta^{mj}(\theta_i + \lambda^m v_i^m) + \omega_i^{mj} > 0)$, $\forall j, m$. Since ω_i^{mj} are iid within each m , we can estimate a separate measurement system for each m to recover unbiased estimates of $(\theta_i + \lambda^m v_i^m)/(1 + (\lambda^m)^2)$ for each i and each m . Finally, under the assumption that v_i^m is independent across i and m , we can use confirmatory factor analysis on our estimates of $(\theta_i + \lambda^m v_i^m)/(1 + (\lambda^m)^2)$ to recover unbiased estimates of θ_i .
- Column (6) instead combines these first-stage estimates of $(\theta_i + \lambda^m v_i^m)/(1 + (\lambda^m)^2)$ using the first factor from an exploratory factor model, i.e. not imposing a single underlying factor.
- Column (7) combines the externally-standardized scores using a confirmatory linear factor model.

Note that checks in columns (3) and (4) are not relevant for socio-emotional problems since there are no binary items here. The checks in columns (5) through (7) are not relevant for socio-emotional problems since we have only one assessment capturing this concept. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. Due to computational intensity, we do not re-estimate the measurement model in 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.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.8: Sensitivity of Estimates to Different Control Variables

	<i>Panel A: Cognitive</i>							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HIM	-0.002 (0.096) p=0.984	0.011 (0.092) p=0.898	0.020 (0.090) p=0.837	0.014 (0.086) p=0.879	-0.012 (0.087) p=0.912	-0.016 (0.095) p=0.868	-0.008 (0.087) p=0.925	-0.029 (0.082) p=0.705
FE+HIM	0.127 (0.085) p=0.131	0.150* (0.085) p=0.073	0.156** (0.083) p=0.048	0.146* (0.078) p=0.056	0.159** (0.079) p=0.038	0.151** (0.075) p=0.040	0.161** (0.073) p=0.025	0.140** (0.068) p=0.036
Difference	0.129** (0.061) p=0.031	0.139** (0.059) p=0.016	0.136** (0.058) p=0.017	0.132*** (0.052) p=0.010	0.171*** (0.051) p=0.001	0.167** (0.070) p=0.016	0.169*** (0.060) p=0.005	0.169*** (0.058) p=0.005
N	1074	1074	1074	1074	1074	1074	1074	1074
Controls								
Age		X	X	X	X		X	X
Gender			X	X	X		X	X
City				X			X	X
Sampling Triplet					X			
Baseline Child Development (factor scores)						X	X	
Baseline Child Development (raw scores)								X
	<i>Panel B: Socio-Emotional Problems</i>							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
HIM	-0.073 (0.146) p=0.613	-0.076 (0.142) p=0.592	-0.089 (0.144) p=0.532	-0.061 (0.149) p=0.684	-0.034 (0.155) p=0.811	-0.040 (0.155) p=0.802	-0.014 (0.155) p=0.933	0.015 (0.145) p=0.914
FE+HIM	-0.180 (0.170) p=0.300	-0.188 (0.169) p=0.274	-0.196 (0.172) p=0.266	-0.157 (0.170) p=0.355	-0.166 (0.166) p=0.327	-0.189 (0.176) p=0.296	-0.155 (0.174) p=0.382	-0.118 (0.168) p=0.497
Difference	-0.106 (0.141) p=0.451	-0.111 (0.140) p=0.427	-0.107 (0.139) p=0.442	-0.097 (0.143) p=0.506	-0.132 (0.143) p=0.342	-0.150 (0.146) p=0.301	-0.141 (0.148) p=0.341	-0.133 (0.146) p=0.368
N	1074	1074	1074	1074	1074	1074	1074	1074
Controls								
Age		X	X	X	X		X	X
Gender			X	X	X		X	X
City				X			X	X
Sampling Triplet					X			
Baseline Child Development (factor scores)						X	X	
Baseline Child Development (raw scores)								X

Note. Table shows impacts on the cognitive and socio-emotional factors controlling for different sets of control variables. Column 1 is for a specification using no controls. Columns 2-5 add age, gender, city and baseline child development controls. Column 6 contains all of these controls and is our main specification shown in Table 3. Column 7 controls for baseline child development using raw scores rather than factor scores estimated using a measurement model. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations) and re-estimating the measurement system on each bootstrap. Clustered standard errors (bootstrapped) in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.9: Impacts on Externally-Standardized Scores in the Extended Sample

	Fluid Reasoning	Expressive Language	School Readiness
HIM only	0.231 (0.426) <i>p</i> =0.573	-0.715 (1.167) <i>p</i> =0.546	0.216 (0.737) <i>p</i> =0.754
HIM+FE	0.939** (0.415) <i>p</i> =0.026	1.898* (1.041) <i>p</i> =0.067	1.185** (0.608) <i>p</i> =0.046
Difference	0.709** (0.332) <i>p</i> =0.032	2.612*** (0.870) <i>p</i> =0.004	0.969 (0.627) <i>p</i> =0.124
N	1837	1833	1835
Control mean	484.78	457.00	45.95
Control SD	6.52	16.89	11.44

Note. Table shows impacts in the extended sample for all three assessments where significant ($p < 0.05$) results were seen in the main analysis sample. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) 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 Section 4.2.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.10: Comparison of Impacts on Factor Scores for Main Analysis and Extended Samples

	Cognitive			Socio-Emotional		
	(1) Main Analysis Sample	(2) Extended Sample	(3) Difference	(4) Main Analysis Sample	(5) Extended Sample	(6) Difference
HIM only	-0.008 (0.087) <i>p</i> =0.925	-0.004 (0.067) <i>p</i> =0.943	0.005 (0.046) <i>p</i> =0.911	-0.014 (0.155) <i>p</i> =0.933	0.047 (0.117) <i>p</i> =0.696	0.061 (0.072) <i>p</i> =0.398
FE+HIM	0.161** (0.073) <i>p</i> =0.025	0.111** (0.056) <i>p</i> =0.044	-0.050 (0.042) <i>p</i> =0.226	-0.155 (0.174) <i>p</i> =0.382	-0.052 (0.134) <i>p</i> =0.694	0.103 (0.084) <i>p</i> =0.222
Difference	0.169*** (0.060) <i>p</i> =0.005	0.114** (0.051) <i>p</i> =0.027	-0.055 (0.036) <i>p</i> =0.141	-0.141 (0.148) <i>p</i> =0.341	-0.099 (0.115) <i>p</i> =0.379	0.042 (0.082) <i>p</i> =0.606
N	1074	1839	1839	1074	1838	1838

Note. Table shows impacts in the main analysis sample (also shown in Table 3) and the extended sample (which includes children below the age of 48 months at endline) for both cognitive and socio-emotional development. For the extended sample, the measurement system is estimated using all children in the extended sample and all items from the measures that were asked to the whole sample. Columns 3 and 6 show the differences in estimates between the samples. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations) and re-estimating the measurement system on each bootstrap. Clustered standard errors (bootstrapped) 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.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.11: Heterogeneity by Age within the Extended Sample

	Cognitive development (1)	Socioemotional problems (2)
(a) HIM x Youngest	-0.023 (0.075) <i>p=0.756</i>	0.106 (0.126) <i>p=0.410</i>
(b) HIM x Middle	-0.031 (0.095) <i>p=0.728</i>	-0.012 (0.178) <i>p=0.943</i>
(c) HIM x Oldest	0.025 (0.086) <i>p=0.777</i>	-0.043 (0.221) <i>p=0.832</i>
(d) HIM+FE x Youngest	0.102 (0.071) <i>p=0.147</i>	-0.056 (0.145) <i>p=0.701</i>
(e) HIM+FE x Middle	0.156 (0.096) <i>p=0.104</i>	-0.294 (0.222) <i>p=0.213</i>
(f) HIM+FE x Oldest	0.151* (0.087) <i>p=0.084</i>	0.027 (0.227) <i>p=0.909</i>
N	1839	1800

Note. Table shows estimates of heterogeneous treatment effects by age in the extended sample (which includes children below the age of 48 months at endline) for both cognitive and socio-emotional development. The measurement system is estimated using all children in the extended sample and all items from the measures that were asked to the whole sample. Two-sided *p*-values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). We re-estimate the measurement system on each bootstrapped sample. Clustered standard errors (bootstrapped) 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 in addition to the heterogeneity variable (terciles of age). Sample is split according to terciles of baseline age. Younger includes children between 18 and 27.7 months at baseline, middle includes children between 27.8 and 32.3 months, older includes children at least 32.4 months.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.12: Impacts on Teachers' Behavior amongst Teachers who were Employed at Baseline

	<i>Teachers' time</i>	<i>Teachers' activities</i>	
	Overtime	Learning	Care
	(1)	(2)	(3)
HIM only	-0.565** (0.272) p=0.041	-0.299** (0.147) p=0.043	-1.018*** (0.275) p=0.002
FE+HIM	0.098 (0.369) p=0.802	0.083 (0.149) p=0.569	-0.564* (0.298) p=0.061
Difference	0.663** (0.316) p=0.037	0.382*** (0.151) p=0.008	0.454 (0.284) p=0.104
N	544	544	543

Note. This table reproduces Table 6 for teachers who were employed in the HI at baseline. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). We re-estimate the measurement system on each bootstrapped sample. Clustered standard errors (bootstrapped) in parentheses. All estimates control for HI-level averages of teachers' learning and care activities, and overtime, measured at baseline, in addition to city effects. Overtime is measured in hours per week. The other variables are factor scores scaled to have a mean of 0 and standard deviation of 1 in the control group (HIM group in the case of TA activities). All factors constructed as described in Section 4.3.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.13: Impacts on Reading Routines within the Home

	(1) Number of books	(2) Minutes reading with mother	(3) Minutes reading with father	(4) Minutes reading by self
HIM	0.207 (0.287) $p=0.476$	-3.201 (13.079) $p=0.803$	-7.248 (7.758) $p=0.363$	1.888 (10.241) $p=0.841$
HIM+FE	0.256 (0.292) $p=0.380$	1.470 (13.968) $p=0.909$	6.585 (9.975) $p=0.520$	-6.276 (9.290) $p=0.490$
Difference	0.050 (0.241) $p=0.833$	4.672 (11.035) $p=0.670$	13.832* (7.890) $p=0.073$	-8.164 (8.440) $p=0.340$
N	1074	1064	835	1074

Note. Table shows impacts on reading routines within the home. In particular: (1) shows impacts on the number of children's story books that the child has access to at home; (2) shows impacts on the time spent reading with their mother in the last 7 days (in minutes); (3) on time spent reading with their father in the last 7 days; (4) on time spent reading by their self in the last 7 days. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. All estimates control for age, gender, city effects, and baseline values of the outcome variables.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.14: Impacts on Anthropometric Outcomes

	Weight for Age Z-Score	Length for Age Z-Score	BMI for Age Z-Score	Weight for Length Z-Score	Acute Malnutrition	Obese	Chronic Malnutrition
HIM	0.033 (0.055) $p=0.530$ $q=0.859$	0.059 (0.040) $p=0.138$ $q=0.43$	-0.012 (0.074) $p=0.872$ $q=0.966$	-0.004 (0.075) $p=0.970$ $q=0.97$	-0.003 (0.014) $p=0.836$ $q=0.966$	0.021 (0.016) $p=0.176$ $q=0.455$	-0.036** (0.016) $p=0.025$ $q=0.106$
HIM+FE	0.083 (0.058) $p=0.144$ $q=0.903$	0.013 (0.034) $p=0.703$ $q=0.459$	0.107 (0.078) $p=0.165$ $q=0.459$	0.118 (0.079) $p=0.130$ $q=0.358$	-0.002 (0.013) $p=0.894$ $q=0.889$	0.028* (0.016) $p=0.079$ $q=0.459$	-0.026 (0.019) $p=0.178$ $q=0.459$
Difference	0.050 (0.057) $p=0.380$ $q=0.833$	-0.046 (0.035) $p=0.191$ $q=0.594$	0.119* (0.071) $p=0.098$ $q=0.387$	0.121* (0.072) $p=0.093$ $q=0.363$	0.001 (0.013) $p=0.925$ $q=0.925$	0.007 (0.015) $p=0.653$ $q=0.918$	0.010 (0.017) $p=0.584$ $q=0.918$
N	1065	1064	1064	1064	1064	1064	1064

Note. Table shows impacts on anthropometric outcomes. In particular, (1)-(4) show impacts on Z-Scores (constructed using WHO's recommended algorithm) of Weight for Age, Length for Age, BMI for Age, and Weight for Length. (5) shows impacts on Acute Malnutrition (or stunting) which is defined by a Length for Age Z-Score of less than -2. (6) shows impacts on obesity which is defined by a Weight for Height Z-Score of more than 2. (7) shows impacts on Chronic Malnutrition which is defined by a Weight for Height Z-Score of less than -2. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) in parentheses. 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 et al. (2019). All estimates control for age, gender, city effects, and baseline values of the outcome variables.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.15: Instrumented Estimates of Impacts of ECERS of Cognitive Development

	<i>First stage</i>	<i>Reduced form</i>	<i>IV</i>
Dep. var.	ECERS	Cognitive Development	Cognitive Development
	(1)	(2)	(3)
HIM	0.056 (0.072) $p=0.439$	0.025 (0.122) $p=0.841$	
HIM+FE	0.160* (0.089) $p=0.073$	0.191** (0.092) $p=0.038$	
ECERS			1.241 (0.767) $p=0.106$
N	726	726	726

Note. This table presents two stage least squares estimates of the local average treatment effect of observed classroom quality (as measured by ECERS) on children's cognitive development under the assumption that the two treatment arms only impact cognitive development through improvements in the ECERS measure. Two-sided p -values estimated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Clustered standard errors (bootstrapped) 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 factors constructed as described in Section 4.3.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C Measurement System Parameters

Table C.1: Measurement Model Parameters: Child Cognitive Development

Assessment	Item (j)	β_j		α_j		g_j	
TVIP	4	1.008	[.663,1.353]	α_j	2.267	[1.871,2.662]	0
TVIP	6	.657	[.375,.94]	α_j	1.635	[1.337,1.934]	0
TVIP	7	1.562	[.893,2.231]	α_j	1.255	[.645,1.866]	.209
TVIP	8	.988	[.062,1.915]	α_j	-.15	[-1.72,1.419]	.544
TVIP	9	.804	[.498,1.109]	α_j	1.618	[1.315,1.921]	0
TVIP	10	.778	[.485,1.072]	α_j	1.245	[.976,1.514]	0
TVIP	12	1.66	[.515,2.805]	α_j	.721	[-.56,2.003]	.599
TVIP	13	1.876	[-.569,4.321]	α_j	-1.121	[-4.24,1.998]	.671
TVIP	14	1.442	[.565,2.319]	α_j	.188	[-.941,1.318]	.443
TVIP	15	1.477	[.082,2.873]	α_j	.054	[-1.762,1.871]	.512
TVIP	16	1.387	[.945,1.83]	α_j	2.306	[1.882,2.729]	0
TVIP	17	3.104	[1.352,4.855]	α_j	.334	[-.616,1.283]	.434
TVIP	18	2.387	[1.041,3.732]	α_j	.6	[-.388,1.589]	.416
TVIP	19	.855	[.534,1.175]	α_j	.986	[.729,1.243]	0
TVIP	20	1.201	[.578,1.823]	α_j	1.248	[.273,2.223]	.32
TVIP	21	1.369	[.555,2.182]	α_j	.043	[-1.059,1.145]	.291
TVIP	22	1.037	[.303,1.771]	α_j	-.264	[-1.562,1.033]	.38
TVIP	23	1.159	[.741,1.577]	α_j	1.44	[1.132,1.748]	0
TVIP	24	2.066	[.349,3.782]	α_j	-1.464	[-3.549,.62]	.505
TVIP	25	1.245	[.788,1.701]	α_j	1.556	[1.23,1.882]	0
TVIP	26	2.027	[.749,3.306]	α_j	.061	[-1.348,1.469]	.606
TVIP	27	2.169	[.657,3.68]	α_j	-1.055	[-2.782,.672]	.546
TVIP	28	.73	[.375,1.085]	α_j	.941	[.676,1.207]	0
TVIP	29	.765	[.431,1.098]	α_j	.34	[.093,.586]	0
TVIP	30	.897	[.535,1.259]	α_j	.606	[.349,.864]	0
TVIP	31	1.257	[.873,1.641]	α_j	-.546	[-.823,-.268]	0
TVIP	32	.443	[.085,.801]	α_j	.838	[.565,1.111]	0
TVIP	33	.962	[.589,1.335]	α_j	.059	[-.205,.324]	0
TVIP	35	3.448	[.834,6.062]	α_j	-1.828	[-4.236,.58]	.536
TVIP	36	.352	[-.001,.706]	α_j	.71	[.432,.989]	0
TVIP	37	.879	[.515,1.242]	α_j	-.366	[-.648,-.085]	0
TVIP	38	.298	[-.024,.62]	α_j	-.252	[-.524,.019]	0
TVIP	39	1.745	[.536,2.954]	α_j	-1.672	[-3.354,.01]	.191
TVIP	40	2.44	[.98,3.9]	α_j	-3.171	[-5.167,-1.174]	.332
TVIP	41	1.14	[-.84,3.12]	α_j	-1.824	[-5.838,2.189]	.443
TVIP	42	1.86	[.016,3.704]	α_j	-3.928	[-7.267,-.589]	.224
TVIP	43	1.132	[.633,1.631]	α_j	.594	[.256,.932]	0
WM14	4	.103	[-.161,.366]	α_j	1.455	[1.188,1.721]	0
WM14	10	.977	[.637,1.316]	α_j	2.219	[1.832,2.606]	0
WM14	14	1.331	[.943,1.72]	α_j	2.237	[1.831,2.644]	0
WM14	15	.906	[.599,1.212]	α_j	1.614	[1.308,1.92]	0
WM14	16	1.023	[.726,1.32]	α_j	.736	[.491,.981]	0
WM14	17	1.135	[.818,1.451]	α_j	.639	[.394,.884]	0
WM14	18	1.553	[1.14,1.967]	α_j	1.727	[1.38,2.075]	0
WM14	19	1.806	[1.389,2.223]	α_j	.228	[-.029,.484]	0
WM14	20	1.439	[1.082,1.795]	α_j	.219	[-.025,.463]	0
WM14	21	1.899	[1.391,2.406]	α_j	-2.458	[-2.941,-1.975]	0
WM14	22	1.4	[1.023,1.778]	α_j	-1.401	[-1.713,-1.089]	0
WM14	23	1.573	[1.127,2.018]	α_j	-1.786	[-2.175,-1.397]	0
WM14	24	1.377	[.986,1.769]	α_j	.012	[-.245,.27]	0
WM17	4	.725	[.395,1.055]	α_j	2.213	[1.838,2.587]	0
WM17	7	1.381	[.99,1.772]	α_j	1.814	[1.458,2.17]	0
WM17	8	1.024	[.709,1.339]	α_j	1.148	[.873,1.422]	0
WM17	9	1.512	[.829,2.194]	α_j	.408	[-.285,1.1]	.22
WM17	10	.756	[.428,1.085]	α_j	.958	[.693,1.223]	0
WM17	11	1.001	[.608,1.394]	α_j	1.538	[1.217,1.858]	0
WM17	12	.935	[.573,1.296]	α_j	1.189	[.903,1.475]	0
WM17	13	.618	[.303,.933]	α_j	-.031	[-.289,.227]	0
WM17	14	.847	[.471,1.222]	α_j	-1.47	[-1.82,-1.12]	0
WM17	15	.798	[.443,1.154]	α_j	-1.195	[-1.515,-.875]	0
Daberon	2	1.644	[.971,2.318]	α_j	2.247	[1.53,2.964]	.271
Daberon	3	1.097	[.77,1.424]	α_j	1.646	[1.328,1.963]	0
Daberon	6	.819	[.522,1.115]	α_j	1.677	[1.367,1.986]	0
Daberon	10	1.322	[.971,1.673]	α_j	1.33	[1.035,1.626]	0
Daberon	11	1.331	[.951,1.711]	α_j	2.147	[1.755,2.54]	0
Daberon	14	3.179	[2.002,4.357]	α_j	1.703	[1.149,2.257]	.231
Daberon	15	3.809	[2.431,5.187]	α_j	2.082	[1.49,2.674]	.175

Daberon	16	3.115	[2.018,4.212]	α_j	1.334	[.799,1.869]	.264	[.146,.428]
Daberon	17	1.877	[1.43,2.323]	α_j	1.291	[.975,1.606]	0	.
Daberon	18	3.131	[1.641,4.621]	α_j	1.38	[.71,2.05]	.401	[.243,.582]
Daberon	19	1.769	[.948,2.589]	α_j	.917	[.27,1.564]	.229	[.066,.556]
Daberon	20	1.802	[1.375,2.229]	α_j	1.069	[.775,1.363]	0	.
Daberon	21	.822	[.55,1.093]	α_j	.892	[.646,1.138]	0	.
Daberon	22	1.088	[.786,1.389]	α_j	-.209	[-.442,.024]	0	.
Daberon	24	1.484	[1.123,1.844]	α_j	.358	[.111,.606]	0	.
Daberon	25	1.27	[.328,2.212]	α_j	1.265	[-.041,2.57]	.429	[.088,.853]
Daberon	26	1.272	[.944,1.601]	α_j	-.344	[-.585,-.102]	0	.
Daberon	27	3.2	[1.411,4.99]	α_j	1.388	[.548,2.229]	.568	[.402,.72]
Daberon	28	1.284	[.952,1.616]	α_j	.652	[.402,.902]	0	.
Daberon	29	1.568	[1.013,2.124]	α_j	-.81	[1.393,-.227]	.139	[.061,.286]
Daberon	30	1.797	[.437,3.157]	α_j	1.311	[.124,2.499]	.554	[.246,.825]
Daberon	32	2.038	[.427,3.649]	α_j	-2.522	[4.78,-.263]	.394	[.298,.5]
Daberon	33	1.454	[.655,2.252]	α_j	1.619	[.752,2.486]	.253	[.029,.793]
Daberon	34	1.541	[.837,2.244]	α_j	-.713	[1.491,.065]	.163	[.061,.369]
Daberon	35	1.963	[.85,3.076]	α_j	-.557	[1.67,.556]	.575	[.443,.697]
Daberon	36	1.228	[.282,2.175]	α_j	-.68	[2.056,.697]	.511	[.332,.688]
Daberon	37	.725	[.432,1.018]	α_j	1.746	[1.433,2.059]	0	.
Daberon	38	.705	[.446,.964]	α_j	-.432	[.658,-.207]	0	.
Daberon	39	1.014	[.246,1.782]	α_j	-1.073	[2.34,.195]	.249	[.101,.494]
Daberon	42	2.071	[.199,3.942]	α_j	-.86	[2.94,1.221]	.768	[.641,.86]
Daberon	43	1.025	[.731,1.319]	α_j	.675	[.432,.917]	0	.
Daberon	44	2.17	[1.222,3.118]	α_j	-1.618	[2.596,-.64]	.13	[.06,.258]
Daberon	46	.787	[.455,1.119]	α_j	2.289	[1.9,2.678]	0	.
Daberon	48	.321	[.018,.625]	α_j	1.962	[1.639,2.285]	0	.
Daberon	49	1.662	[1.238,2.087]	α_j	1.833	[1.468,2.198]	0	.
Daberon	50	.386	[.162,.61]	α_j	.421	[.204,.638]	0	.
Daberon	51	.833	[.523,1.143]	α_j	1.907	[1.571,2.244]	0	.
Daberon	52	.779	[.491,1.067]	α_j	1.559	[1.264,1.855]	0	.
Daberon	53	.753	[.467,1.039]	α_j	1.569	[1.274,1.865]	0	.
Daberon	54	.998	[.688,1.308]	α_j	1.524	[1.223,1.824]	0	.
Daberon	55	.965	[.633,1.297]	α_j	2.101	[1.733,2.47]	0	.
Daberon	56	1.276	[.574,1.978]	α_j	.207	[.672,1.085]	.283	[.095,.598]
Daberon	57	.699	[.435,.964]	α_j	-.656	[.888,-.423]	0	.
Daberon	58	-.011	[.225,.202]	α_j	.432	[.218,.645]	0	.
Daberon	59	.497	[.259,.734]	α_j	.752	[.521,.982]	0	.
Daberon	60	.616	[.346,.885]	α_j	-.976	[1.221,-.73]	0	.
Daberon	62	.663	[.38,.946]	α_j	1.638	[1.339,1.937]	0	.
Daberon	63	1.131	[.825,1.437]	α_j	-.031	[.263,.202]	0	.
Daberon	64	.901	[.581,1.221]	α_j	-1.42	[1.71,-1.13]	0	.
Daberon	65	1.503	[1.14,1.866]	α_j	-.343	[.593,-.094]	0	.
Daberon	66	1.91	[1.481,2.34]	α_j	-.205	[.466,.055]	0	.
Daberon	67	1.313	[.977,1.648]	α_j	-.424	[.669,-.178]	0	.
Daberon	68	1.157	[.799,1.514]	α_j	-1.603	[1.923,-1.282]	0	.
Daberon	69	1.177	[.837,1.517]	α_j	-1.242	[1.528,-.955]	0	.
Daberon	70	.844	[.55,1.138]	α_j	-1.035	[1.292,-.777]	0	.
PTT		.914	[.714,1.115]	α_{j1}	3.136	[2.665,3.606]		
PTT				α_{j2}	2.368	[2.016,2.72]		
PTT				α_{j3}	1.836	[1.54,2.133]		
PTT				α_{j4}	1.331	[1.071,1.59]		
PTT				α_{j5}	.847	[.61,1.084]		
PTT				α_{j6}	.523	[.295,.751]		
PTT				α_{j7}	.248	[.024,.471]		
PTT				α_{j8}	-.128	[.351,.095]		
PTT				α_{j9}	-.511	[.74,-.283]		
PTT				α_{j10}	-.848	[1.087,-.609]		
PTT				α_{j11}	-1.349	[1.613,-1.084]		
PTT				α_{j12}	-1.763	[2.057,-1.468]		
PTT				α_{j13}	-2.083	[2.407,-1.759]		
PTT				α_{j14}	-2.408	[2.768,-2.048]		
PTT				α_{j15}	-2.948	[3.382,-2.513]		
PTT				α_{j16}	-3.515	[4.057,-2.974]		
WM12	a	1.048	[.837,1.26]	α_{j1}	2.92	[2.488,3.352]		
WM12	a			α_{j2}	2.077	[1.75,2.403]		
WM12	a			α_{j3}	1.299	[1.035,1.562]		
WM12	a			α_{j4}	.648	[.414,.882]		
WM12	a			α_{j5}	.008	[.216,.232]		
WM12	a			α_{j6}	-.721	[.957,-.486]		
WM12	a			α_{j7}	-1.409	[1.676,-1.142]		

WM12	a			α_{j8}	-2.082	[-2.402,-1.763]
WM12	a			α_{j9}	-2.55	[-2.922,-2.178]
WM12	a			α_{j10}	-3.332	[-3.83,-2.834]
WM12	c	1.322	[1.098,1.546]	α_{j1}	2.824	[2.408,3.239]
WM12	c			α_{j2}	1.946	[1.626,2.265]
WM12	c			α_{j3}	1.077	[.817,1.336]
WM12	c			α_{j4}	.408	[.172,.644]
WM12	c			α_{j5}	-.294	[-.527,-.062]
WM12	c			α_{j6}	-.954	[-1.203,-.705]
WM12	c			α_{j7}	-1.948	[-2.256,-1.64]
WM12	c			α_{j8}	-2.553	[-2.918,-2.188]
WM12	c			α_{j9}	-3.016	[-3.44,-2.592]
WM12	c			α_{j10}	-3.632	[-4.16,-3.104]

Notes: Table presents estimated parameters for IRT measurement model of children's cognitive development alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the Daberon, TVIP, WM-14 and WM-17 are binary and we model them using 3-parameter model with the guessing parameter described in equation (4.1). The PTT, WM-12(a) and WM-12(b) are ordinal and we model them using the graded response model described in equation (4.2).

Table C.2: Measurement Model Parameters: Child Socioemotional Problems

Item (j)	β_j	α_{j1}	α_{j2}
1	1.08 [-.688,1.472]	-1.64 [-1.999,-1.286]	-4.60 [-5.486,-3.711]
2	0.62 [-.02,1.254]	-3.36 [-4.001,-2.715]	
3	0.44 [-.169,.717]	-0.90 [-1.146,-.659]	-2.64 [-3.053,-2.219]
4	0.06 [-.191,.314]	2.76 [2.322,3.205]	0.69 [.469,.912]
5	1.13 [-.794,1.473]	-0.55 [-.827,-.276]	-2.48 [-2.903,-2.065]
6	-0.03 [-.26,.196]	-0.06 [-.27,.148]	-1.41 [-1.677,-1.151]
7	1.30 [-.91,1.699]	-0.81 [-1.117,-.509]	-3.98 [-4.67,-3.293]
8	0.91 [-.586,1.236]	-0.88 [-1.149,-.603]	-3.40 [-3.961,-2.844]
9	1.40 [-.718,2.08]	-3.34 [-4.163,-2.512]	
10	0.96 [-.574,1.342]	-1.79 [-2.152,-1.425]	-4.20 [-4.971,-3.428]
11	1.32 [-.922,1.715]	0.33 [.044,.607]	-4.30 [-5.075,-3.531]
12	-0.16 [-.41,.085]	2.94 [2.462,3.413]	0.42 [.21,.639]
14	0.49 [-.198,.778]	-1.07 [-1.324,-.815]	-3.37 [-3.944,-2.805]
15	0.79 [-.326,1.251]	-2.55 [-3.016,-2.082]	-4.35 [-5.225,-3.479]
16	0.72 [-.232,1.198]	-2.71 [-3.196,-2.215]	-3.78 [-4.482,-3.072]
17	1.89 [1.001,2.777]	-4.33 [-5.537,-3.126]	
19	0.87 [-.581,1.153]	0.47 [.219,.713]	-1.98 [-2.326,-1.641]
21	0.55 [-.273,.831]	-0.79 [-1.036,-.551]	-2.14 [-2.493,-1.796]
22	0.96 [-.303,1.623]	-3.52 [-4.266,-2.768]	
23	1.04 [-.628,1.447]	-2.04 [-2.444,-1.632]	-4.27 [-5.068,-3.48]
24	0.76 [-.435,1.085]	-1.19 [-1.474,-.908]	-3.53 [-4.129,-2.935]
25	1.43 [-.888,1.978]	-2.74 [-3.327,-2.148]	-5.72 [-7.028,-4.405]
26	1.09 [-.486,1.686]	-2.99 [-3.667,-2.318]	-5.15 [-6.429,-3.864]
27	1.25 [-.734,1.773]	-2.66 [-3.205,-2.106]	
28	1.37 [-.823,1.925]	-2.75 [-3.34,-2.161]	
29	0.79 [-.403,1.175]	-1.94 [-2.306,-1.575]	-4.54 [-5.462,-3.612]
30	0.11 [-.12,.347]	-0.15 [-.355,.065]	-2.09 [-2.427,-1.759]
31	0.17 [-.25,.583]	-2.07 [-2.439,-1.697]	-2.93 [-3.453,-2.4]
32	1.65 [1.042,2.267]	-2.95 [-3.625,-2.279]	-6.42 [-8.044,-4.804]
33	0.62 [-.345,.889]	-0.50 [-.734,-.263]	-2.37 [-2.749,-1.993]
34	1.23 [-.861,1.595]	-0.41 [-.686,-.13]	-4.00 [-4.689,-3.319]
35	0.84 [-.496,1.191]	-1.4 [-1.748,-1.127]	-3.2 [-3.758,-2.691]

Notes: Table presents estimated parameters for IRT measurement model of children's socioemotional problems alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items are ordinal (taking the values 0, 5 or 10) and we model them using the graded response model described in equation (4.2).

Table C.3: Measurement Model Parameters: Baseline Child Development

<i>Panel A: ASQ-Communication</i>					<i>Panel B: ASQ-Gross Motor</i>				
Item (j)	β_j		α_j		Item (j)	β_j		α_j	
Item 1	0.89	[-.284,2.063]	2.61	[1.502,3.725]	Item 1	0.91	[.384,1.427]	1.71	[1.28,2.132]
Item 2	2.04	[-.254,4.34]	3.90	[1.097,6.694]	Item 2	1.65	[.929,2.367]	2.60	[1.927,3.274]
Item 3	0.73	[.148,1.305]	0.64	[.282,1.006]	Item 3	1.66	[.744,2.568]	3.10	[2.157,4.05]
Item 4	0.96	[.467,1.449]	1.73	[1.343,2.108]	Item 4	0.90	[.284,1.52]	1.30	[.837,1.757]
Item 5	1.14	[.609,1.667]	0.65	[.36, .939]	Item 5	0.82	[.431,1.211]	1.22	[.925,1.519]
Item 6	2.70	[.444,4.955]	1.26	[.393,2.116]	Item 6	1.63	[.737,2.514]	3.81	[2.745,4.873]
Item 7	1.32	[-.276,2.914]	3.58	[1.691,5.473]	Item 7	1.45	[.267,2.629]	2.28	[1.224,3.331]
Item 8	2.06	[-.015,4.127]	3.96	[1.524,6.388]	Item 8	0.56	[.244, .876]	0.30	[.071, .526]
Item 9	0.90	[.462,1.339]	1.34	[1.014,1.656]	Item 9	1.65	[.434,2.866]	1.75	[.874,2.621]
Item 10	0.91	[.084,1.739]	1.48	[.855,2.112]	Item 10	0.86	[.152,1.563]	0.33	[-.14, .803]
Item 11	1.23	[.622,1.827]	2.35	[1.808,2.888]	Item 11	1.36	[.831,1.897]	1.38	[.999,1.765]
Item 12	0.66	[-.111,1.422]	2.23	[1.621,2.83]	Item 12	1.73	[1.052,2.399]	0.86	[.491,1.237]
Item 13	0.59	[.171,1.014]	1.15	[.837,1.467]					
Item 14	0.47	[-.177,1.11]	0.64	[.188,1.09]					
<i>Panel C: ASQ-Fine Motor</i>					<i>Panel D: ASQ-Problem Solving</i>				
Item (j)	β_j		α_j		Item (j)	β_j		α_j	
Item 1	0.54	[.176, .911]	-1.09	[-1.392, -.781]	Item 1	0.94	[.221,1.657]	-1.00	[-1.566, -.442]
Item 2	1.28	[.746,1.816]	-1.46	[-1.881, -1.041]	Item 2	0.72	[.386,1.046]	0.71	[.46, .96]
Item 3	0.67	[.365, .974]	0.73	[.486, .982]	Item 3	0.96	[.495,1.429]	2.09	[1.67,2.515]
Item 4	0.12	[-.263, .509]	1.67	[1.332,2.001]	Item 4	5.56	[-1.128,12.242]	-1.26	[-2.827, .317]
Item 5	0.75	[.022,1.479]	-1.73	[-2.362, -1.098]	Item 5	-0.19	[-.636, .262]	1.94	[1.574,2.312]
Item 6	0.52	[.241, .799]	0.40	[.171, .626]	Item 6	0.57	[-.011,1.155]	0.70	[.238,1.159]
Item 7	4.41	[.44,8.377]	1.17	[-.133,2.47]	Item 7	0.38	[-.06, .819]	0.98	[.624,1.344]
Item 8	0.72	[.418,1.024]	0.39	[.152, .628]	Item 8	0.83	[.081,1.577]	1.50	[.897,2.099]
Item 9	5.11	[2.138,8.074]	1.03	[.141,1.921]	Item 9	0.54	[.164, .91]	1.10	[.797,1.406]
Item 10	1.65	[1.141,2.159]	1.08	[.717,1.438]	Item 10	0.95	[.424,1.468]	2.54	[2.033,3.046]
Item 11	3.99	[2.028,5.949]	1.12	[.34,1.899]	Item 11	2.35	[1.455,3.248]	0.96	[.499,1.41]
Item 12	0.71	[.05,1.366]	-1.28	[-1.854, -.715]	Item 12	0.65	[.078,1.225]	0.38	[-.057, .815]
<i>Panel E: ASQ-Socio Individual</i>					<i>Panel F: ASQ-Socio Emotional</i>				
Item (j)	β_j		α_j		Item (j)	β_j		α_j	
Item 1	0.56	[.221, .906]	-0.06	[-.281, .168]	Item 1	0.01	[-.278, .303]	1.13	[.89,1.376]
Item 2	0.66	[.305,1.012]	0.19	[-.038, .424]	Item 2	1.57	[.967,2.167]	-2.79	[-3.403, -2.166]
Item 3	0.42	[.021, .824]	1.59	[1.295,1.893]	Item 3	-0.27	[-.786, .239]	2.79	[2.334,3.253]
Item 4	1.10	[.062,2.129]	0.93	[.378,1.472]	Item 4	0.70	[.355,1.037]	-1.40	[-1.697, -1.108]
Item 5	2.28	[1.3,564]	3.48	[2.17,4.782]	Item 5	1.47	[.809,2.138]	-3.33	[-4.106, -2.557]
Item 6	1.53	[.856,2.211]	1.53	[1.078,1.983]	Item 6	1.42	[.708,2.13]	-3.72	[-4.612, -2.82]
Item 7	0.12	[-.345, .576]	2.15	[1.807,2.494]	Item 7	0.87	[.388,1.351]	-2.67	[-3.172, -2.172]
Item 8	0.79	[.196,1.393]	2.39	[1.842,2.944]	Item 8	1.16	[.663,1.652]	-2.34	[-2.831, -1.857]
Item 9	1.38	[.782,1.975]	0.43	[.132, .727]	Item 9	-0.95	[-2.493, .602]	1.44	[.185,2.693]
Item 10	0.91	[.361,1.459]	1.48	[1.092,1.875]	Item 10	0.64	[.326, .948]	-0.57	[-.813, -.322]
					Item 11	1.32	[.755,1.878]	-2.70	[-3.285, -2.116]
					Item 12	1.57	[1.033,2.11]	-1.16	[-1.534, -.787]
					Item 13	0.75	[.419,1.075]	0.07	[-.17, .316]
					Item 14	0.80	[.416,1.175]	-1.78	[-2.121, -1.433]
					Item 15	0.57	[.255, .875]	-1.11	[-1.377, -.852]
					Item 16	1.39	[-1.114,3.898]	-3.20	[-5.873, -.53]
					Item 17	1.14	[.736,1.538]	-0.80	[-1.087, -.511]
					Item 18	0.31	[-1.206,1.817]	1.57	[.485,2.655]
					Item 19	0.87	[.508,1.235]	0.93	[.658,1.207]
					Item 20	1.06	[.679,1.435]	0.22	[-.043, .474]
					Item 21	1.07	[.688,1.444]	-0.88	[-1.161, -.589]
					Item 22	0.51	[.171, .857]	-1.55	[-1.859, -1.249]
					Item 23	0.09	[-.229, .41]	1.39	[1.115,1.657]
					Item 24	3.41	[-.907,7.718]	-3.94	[-8.061, .175]
					Item 25	1.41	[.941,1.875]	-1.08	[-1.411, -.74]
					Item 26	2.60	[-.703,5.906]	-0.86	[-2.487, .763]
					Item 27	2.08	[-.336,4.503]	0.06	[-1.126,1.25]
					Item 28	0.09	[-.322, .498]	2.18	[1.832,2.525]
					Item 29	1.72	[1.074,2.36]	-2.44	[-3.037, -1.849]
					Item 31	1.84	[-.249,3.92]	-1.65	[-3.246, -.047]
					Item 32	0.27	[-1.31,1.845]	-2.02	[-3.351, -.687]
					Item 33	1.95	[1.078,2.82]	-4.00	[-5.092, -2.91]

Panel G: MacArthur-Bates - Younger Kids

Word	β_j	α_j
Afuera	3.5 [1.298,5.679]	3.53 [1.485,5.568]
Aquí	1.81 [.521,3.092]	2.71 [1.451,3.958]
Besar	0.98 [.24,1.724]	1.36 [.694,2.025]
Bigote	2.40 [1.129,3.662]	-0.89 [-1.765,-.007]
Bonita/linda	3.73 [1.121,6.346]	4.34 [1.685,6.984]
Brazo	1.34 [.415,2.255]	1.79 [.962,2.622]
Buenas noches	1.34 [.481,2.189]	1.32 [.598,2.05]
Bus	2.65 [.47,4.839]	4.25 [1.82,6.672]
Caer(se)	2.48 [.462,4.496]	4.07 [1.84,6.306]
Caliente	3.76 [.593,6.921]	5.40 [1.763,9.042]
Calle	1.61 [.47,2.754]	2.43 [1.333,3.518]
Camisa	1.34 [.397,2.281]	1.90 [1.038,2.76]
Cansado	2.11 [.826,3.402]	2.24 [1.095,3.376]
Carne	2.84 [.749,4.939]	3.94 [1.737,6.144]
Cómo	3.97 [1.745,6.194]	2.20 [.631,3.758]
Comprar	2.6 [1.187,3.993]	1.68 [.596,2.763]
Culebra/serpiente	1.53 [.686,2.371]	-0.10 [-.735,-.546]
Dónde	1.94 [.755,3.128]	2.01 [.983,3.044]
En la mañana	1.54 [.688,2.384]	-0.40 [-1.051,-.253]
Entonces	2.54 [1.208,3.875]	0.71 [-.186,1.608]
Escoba	2.59 [.823,4.351]	3.32 [1.569,5.065]
Estar	2.20 [1.026,3.38]	1.01 [.151,1.873]
Falda	2.49 [1.204,3.771]	0.50 [-.363,1.367]
Fiesta	2.22 [.945,3.497]	1.95 [.875,3.014]
Flor	2.15 [.82,3.486]	2.39 [1.184,3.589]
Fósforos	1.37 [.584,2.148]	-0.53 [-1.158,-.097]
Ganar	1.95 [.9,2.99]	0.85 [.064,1.627]
Gato	2.00 [.173,3.826]	4.22 [2.012,6.417]
Grande	3.33 [1.288,5.363]	3.23 [1.374,5.076]
Haber (hay)	1.62 [.655,2.579]	1.35 [.558,2.141]
Hacer	2.24 [.91,3.575]	2.19 [1.036,3.348]
Hoy	2.68 [1.248,4.113]	1.49 [.436,2.552]
Huevo	4.31 [-.166,8.793]	6.85 [1.176,12.525]
Iglesia	2.29 [1.103,3.47]	0.47 [-.349,1.287]
Jabón	5.74 [-.043,11.522]	8.10 [.975,15.23]
Jugar	6.73 [-.844,14.304]	9.93 [.045,19.804]
Lavamanos	3.91 [1.171,6.657]	4.50 [1.702,7.303]
Leche	2.70 [.478,4.924]	4.29 [1.818,6.766]
Libro	2.53 [.907,4.152]	2.94 [1.41,4.468]
Llover	1.82 [.692,2.948]	1.95 [.961,2.928]
Madrina	2.27 [1.085,3.454]	-0.37 [-1.175,-.431]
Malo	2.61 [1.114,4.111]	2.29 [1.02,3.567]
Manguera	2.86 [1.363,4.354]	-1.10 [-2.114,-.093]
Mirar	2.78 [.712,4.854]	3.88 [1.713,6.054]
Más	2.16 [.394,3.921]	3.77 [1.834,5.703]
No hay	3.22 [1.17,5.27]	3.49 [1.53,5.445]
Nuestro	1.58 [.715,2.446]	-0.72 [-1.407,-.037]
Nuevo	2.58 [1.232,3.929]	1.13 [.166,2.098]
Oír	2.24 [.932,3.55]	2.07 [.956,3.187]
Olla	2.94 [1.124,4.747]	3.09 [1.407,4.775]
Pantalón	4.6 [.917,8.376]	5.75 [1.665,9.827]
Papas	2.2 [.847,3.485]	2.27 [1.101,3.435]
Pato	1.9 [.898,2.929]	0.49 [-.244,1.233]
Periódico	2.21 [1.071,3.344]	-0.75 [-1.566,-.069]
Plátano/banano	2.17 [.895,3.45]	2.03 [.944,3.12]
Por favor	2.77 [1.007,4.542]	3.13 [1.45,4.806]
Prender	1.92 [.784,3.057]	1.78 [.825,2.741]
Puerta	3.52 [.708,6.323]	4.85 [1.758,7.94]
Quién	3.40 [1.475,5.316]	2.49 [.964,4.02]
Quiquiriquí	1.09 [.393,1.779]	0.24 [-.328,-.799]
Rana	1.10 [.408,1.787]	0.30 [-.265,-.871]
Roto	1.73 [.7,2.768]	1.59 [.721,2.457]
Saber	2.5 [1.179,3.746]	0.89 [-.011,1.796]

Panel H: MacArthur-Bates - Older Kids

Word	β_j	α_j
abierto	1.06 [.627,1.493]	1.96 [1.535,2.378]
adelante	2.128 [1.475,2.782]	2.09 [1.539,2.633]
ambulancia	1.26 [.858,1.654]	0.15 [-.158,.451]
aquel	1.00 [.652,1.344]	-0.14 [-.426,-.139]
arreglar	1.75 [1.196,2.313]	1.89 [1.413,2.372]
atrás	1.81 [1.149,2.467]	3.08 [2.362,3.789]
ayer	1.29 [.869,1.707]	0.85 [.521,1.186]
barba	1.21 [.823,1.596]	-0.69 [-1.004,-.375]
biblioteca (pública)	1.78 [1.261,2.296]	-1.29 [-1.703,-.883]
bolsa	1.48 [.873,2.085]	3.14 [2.437,3.836]
caber	1.44 [1.003,1.882]	0.35 [.026,-.678]
cada	1.58 [1.114,2.042]	-0.15 [-.483,-.186]
candado	1.23 [.838,1.625]	0.20 [-.104,503]
cesta o canasta	2.18 [1.557,2.804]	1.02 [.588,1.457]
clínica o hospital	1.65 [1.167,2.135]	0.34 [-.008,.684]
computadora	1.57 [1.022,2.116]	2.31 [1.782,2.833]
contra	1.56 [1.094,2.031]	-1.16 [-1.532,-.778]
pelocorto	1.00 [.583,1.412]	1.84 [1.439,2.237]
cuadrado	1.57 [1.081,2.056]	1.19 [.809,1.576]
cuál	1.30 [.837,1.769]	1.84 [1.414,2.27]
cueva	2.18 [1.571,2.793]	-1.06 [-1.492,-.627]
dañado	1.54 [.913,2.171]	3.25 [2.513,3.985]
descansar	2.39 [1.694,3.094]	1.69 [1.165,2.208]
después	1.92 [1.349,2.497]	1.39 [.952,1.829]
dinosaurio	1.33 [.909,1.754]	0.68 [.357,1.012]
echar	1.64 [1.155,2.116]	0.36 [.011,-.701]
empujar	1.84 [1.209,2.461]	2.63 [2.019,3.249]
enfermo	1.77 [1.134,2.396]	2.89 [2.229,3.556]
escalera	1.81 [1.106,2.506]	3.48 [2.656,4.308]
estantería o armario	1.61 [1.142,2.08]	-0.31 [-.651,-.029]
fábrica	2.66 [1.907,3.413]	-1.68 [-2.222,-1.133]
faltar	2.37 [1.708,3.026]	0.50 [.076,-.928]
figura	2.92 [2.077,3.756]	1.59 [1.02,2.16]
flecha	1.96 [1.406,2.503]	-0.11 [-.478,-.266]
garganta	1.75 [1.23,2.273]	1.00 [.61,1.385]
grupo	2.85 [2.07,3.634]	-0.01 [-.483,455]
hasta	2.25 [1.623,2.87]	0.25 [-.156,655]
herramienta	1.72 [1.225,2.215]	-0.05 [-.394,302]
horno	1.66 [1.181,2.143]	-0.15 [-.492,-.194]
idea	2.18 [1.578,2.786]	-0.20 [-.593,-.2]
igual	2.43 [1.736,3.125]	1.25 [.771,1.731]
insecto	1.36 [.943,1.773]	-0.26 [-.576,-.054]
jalar	1.31 [.872,1.737]	1.15 [.796,1.506]
juntar	1.75 [1.228,2.264]	0.93 [.546,1.311]
lado	2.21 [1.53,2.88]	2.16 [1.594,2.727]
lastimar	2.08 [1.458,2.706]	1.70 [1.21,2.187]
letra	2.44 [1.717,3.166]	1.95 [1.392,2.515]
línea	2.39 [1.716,3.067]	0.95 [.5,1.405]
lugar	2.83 [2.009,3.655]	1.78 [1.194,2.358]
manejar	1.34 [.888,1.782]	1.37 [.989,1.742]
mecánico	1.66 [1.167,2.157]	-1.47 [-1.88,-1.05]
medir	2.85 [2.064,3.625]	-0.48 [-.954,-.005]
meter	1.70 [1.143,2.247]	1.98 [1.496,2.47]
mis	1.15 [.724,1.568]	1.56 [1.178,1.936]
mismo	1.69 [1.19,2.183]	0.59 [.233,-.949]
montaña	2.08 [1.483,2.679]	0.97 [.551,1.391]
mover	2.17 [1.471,2.876]	2.70 [2.039,3.361]
mueble	1.35 [.867,1.833]	1.98 [1.532,2.436]
muy	1.67 [1.144,2.198]	1.60 [1.167,2.036]
necesitar	2.23 [1.605,2.857]	0.60 [.184,1.015]
nido	2.02 [1.453,2.577]	-0.49 [-.875,-.105]
nosotros	2.51 [1.764,3.256]	2.05 [1.469,2.634]
oficina	1.99 [1.431,2.555]	-0.89 [-1.294,-.49]

Saltar	1.9	[.586,3.172]	2.61	[1.375,3.84]	oscuro	1.55	[1.012,2.089]	2.23	[1.716,2.737]
Sentar(se)	1.6	[.134,2.979]	3.78	[2.025,5.543]	parecer	3.13	[2.266,3.99]	-0.65	[-1.162,-.136]
Sol	2.7	[1.063,4.378]	2.78	[1.278,4.278]	peligroso	2.76	[1.994,3.528]	0.67	[.192,1.146]
Ésta	3.3	[1.235,5.34]	3.37	[1.458,5.274]	pequeño	2.67	[1.757,3.583]	3.69	[2.724,4.662]
Sucio	1.9	[.335,3.499]	3.55	[1.819,5.281]	pera	1.80	[1.236,2.368]	1.83	[1.352,2.304]
Suya	1.3	[.376,2.176]	1.76	[.95,2.574]	perder	2.02	[1.44,2.606]	0.93	[.521,1.343]
Tambor	2.4	[1.157,3.606]	0.2	[-.625,1.022]	perfecto	2.59	[1.866,3.308]	-1.27	[-1.766,-.779]
Televisión	1.7	[.451,3.031]	2.82	[1.527,4.103]	perseguir	2.40	[1.733,3.059]	0.29	[-.135,.71]
Tetero	1.2	[.122,2.289]	2.97	[1.747,4.191]	persona	2.49	[1.738,3.235]	2.24	[1.635,2.853]
Tigre	1.4	[.624,2.27]	0.49	[-.151,1.134]	pesado	1.61	[1.068,2.148]	2.09	[1.599,2.587]
Timbre	1.8	[.841,2.747]	0.15	[-.547,.845]	pintor	1.67	[1.189,2.151]	-0.44	[-.788,-.091]
Tomate	1.7	[.692,2.754]	1.58	[.72,2.448]	plástico	2.57	[1.856,3.291]	0.74	[.278,1.198]
Tutu	1.1	[.394,1.768]	0.3	[-.264,.866]	por	1.54	[1.056,2.025]	1.25	[.867,1.639]
Vaca	1.2	[.233,2.234]	2.46	[1.446,3.481]	pulsera	1.62	[1.125,2.119]	1.14	[.756,1.526]
Vámonos	1.33	[.076,2.588]	3.58	[2.002,5.157]	puntilla	1.64	[1.163,2.112]	-0.27	[-.613,.07]
					quedar	1.62	[1.133,2.108]	0.83	[.468,1.194]
					raqueta	1.38	[.957,1.794]	-0.34	[-.657,-.021]
					raro	2.61	[1.895,3.326]	0.08	[-.362,.524]
					regresar	2.39	[1.732,3.054]	0.17	[-.254,.586]
					río	2.32	[1.615,3.017]	2.15	[1.577,2.729]
					saber	2.72	[1.935,3.498]	1.55	[1.005,2.089]
					salvar	2.61	[1.896,3.325]	-0.44	[-.884,.011]
					sembrar	2.36	[1.715,3.013]	-0.32	[-.739,.096]
					semilla	2.21	[1.602,2.825]	-0.46	[-.864,-.054]
					sobre (la silla)	1.79	[1.241,2.341]	1.52	[1.082,1.959]
					sus	1.77	[1.231,2.313]	1.41	[.985,1.835]
					suyos	1.23	[.789,1.678]	1.67	[1.272,2.072]
					también	1.76	[1.175,2.339]	2.25	[1.711,2.78]
					ti	1.01	[.631,1.39]	1.17	[.835,1.494]
					tigre	1.64	[1.098,2.173]	1.95	[1.476,2.428]
					torre	2.47	[1.775,3.167]	0.95	[.486,1.408]
					tractor	1.59	[1.117,2.053]	-1.07	[-1.438,-.694]
					tranquilo	2.77	[2.3,5.46]	0.81	[.319,1.291]
					vainilla	2.037	[1.466,2.607]	-1	[-1.36,-.541]
					vender	2.627	[1.9,3.353]	0.47	[.012,.918]
					verdura	1.588	[1.092,2.083]	1.29	[.899,1.687]
					vidrio	1.794	[1.237,2.352]	1.71	[1.251,2.171]

Notes: Table presents estimated parameters for IRT measurement model of children's development measured at baseline alongside 95% confidence intervals in brackets. We estimate separate measurement models for each baseline measure and present each in a separate panel. Parameters are estimated only using observations in the control group. All items are binary. Since they are answered by parental report there is unlikely to be scope for guessing. Therefore we model all items using the IRT model described in equation (4.1) with the guessing parameter set to 0.

Table C.4: Measurement Model Parameters: Teacher Learning Activities

Item (j)	β_j										α_{j1}										α_{j2}										α_{j3}										α_{j4}																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																	
<i>Number of times did last week</i>																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
Read stories	1.22	[.836, 1.601]	2.39	[1.922, 2.855]	1.72	[1.32, 2.111]	0.76	[.421, 1.09]	0.5	[.17, .82]																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																</

Notes: Table presents estimated parameters for IRT measurement model of teachers' learning activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using an IRT model with the guessing parameter set to zero as described in equation (4.1). Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2).

Table C.5: Measurement Model Parameters: Teacher Care Activities

Item (j)	β_j	α_{j1}	α_{j2}	α_{j3}	α_{j4}
Number of times did last week					
Hygiene routines and care, e.g. changing nappies, brush teeth, wash hands	0.98 [.681,1.276]	5.34 [3.912,6.774]	4.93 [3.741,6.11]	4.63 [3.587,5.667]	3.77
Supply medicines / remedies	0.98 [.681,1.276]	-2.37 [-2.826,-1.918]	-2.86 [-3.382,-2.33]	-3.26 [-3.864,-2.656]	
Pamper children	0.98 [.681,1.276]	4.94 [3.749,6.124]	4.4 [3.457,5.345]	4.21 [3.335,5.077]	3.45
Watch TV with children	0.98 [.681,1.276]	-1.97 [-2.37,-1.562]	-2.63 [-3.119,-2.142]	-3.25 [-3.85,-2.647]	
Describe as being a main duty					
Teach and support children during practices on personal hygiene.	0.98 [.681,1.276]	1.59 [1.218,1.951]			
Support during meals.	0.98 [.681,1.276]	1.78 [1.397,2.169]			

Notes: Table presents estimated parameters for IRT measurement model of teachers' caring activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2). To aid convergence given the small number of items we restrict the discrimination parameter to be equal across items.

Table C.6: Measurement Model Parameters: TA Learning Activities

Item (j)	β_j	α_{j1}	α_{j2}	α_{j3}	α_{j4}					
Number of times did last week										
Read stories	0.51	[136, 883]	0.53	[142, 915]	0.14	[-239, 511]	-0.57	[-956, -183]	-0.8	[-1.196, -398]
Tell stories	0.92	[469, 1.368]	0.39	[-026, 812]	-0.36	[-776, 061]	-0.91	[-1.36, -467]	-1.16	[-1.627, -692]
Conversation	2.34	[1.243, 3.445]	6.49	[4.168, 8.815]	5.29	[3.429, 7.159]	5.02	[3.24, 6.797]	3.07	[1.851, 4.282]
Sing	1.75	[856, 2.642]	5.53	[3.601, 7.467]	5.04	[3.347, 6.741]	2.77	[1.762, 3.767]		
Dance	1.02	[514, 1.526]	2.09	[1.471, 2.698]	1.8	[1.232, 2.362]	0.86	[4.1.316]	0.41	[-0.28, 839]
Watch a video or educational programme on TV.	0.21	[-157, 578]	-0.12	[-478, 239]	-2	[-2.542, -1.452]	-2.83	[-3.598, -2.064]	-2.99	[-3.818, -2.17]
isit other places in the community	0.21	[-333, 753]	-1.78	[-2.29, -1.271]						
Free play within the nursery premises	0.62	[21, 1.032]	1.84	[1.31, 2.367]	1.07	[635, 1.502]	0.59	[192, 995]	0.3	[-093, 687]
Free play in the recreation area	0.39	[018, 758]	0.69	[303, 1.074]	0.13	[-233, 498]	-0.37	[-739, 003]	-0.69	[-1.079, -307]
Physical activities such as running, jumping	0.7	[302, 1.096]	1.71	[1.199, 2.222]	0.81	[395, 1.234]	0.18	[-217, 569]	-0.19	[-583, 202]
Group learning activities.	1.4	[822, 1.969]	1.74	[1.14, 2.346]	1.07	[536, 1.598]	0.62	[117, 1.113]	0.33	[-151, 819]
Individual learning activities.	1.3	[732, 1.874]	0.76	[27, 1.248]	0.4	[-074, 869]	0.1	[-362, 568]	-0.11	[-57, 357]
Teach colours	1.59	[905, 2.265]	0.02	[-484, 528]	-0.57	[-1.087, -046]	-0.93	[-1.476, -39]	-1.19	[-1.752, -623]
Teach numbers	1.06	[527, 1.589]	-0.2	[-632, 233]	-0.77	[-1.225, -312]	-1.05	[-1.531, -573]	-1.21	[-1.703, -713]
Teach letter of the alphabet	1.14	[467, 1.805]	-1.69	[-2.294, -1.087]	-2.25	[-2.952, -1.553]	-2.54	[-3.299, -1.775]	-2.65	[-3.434, -1.857]
Teach forms and shapes	1.13	[624, 1.632]	0.4	[-044, 844]	-0.53	[-976, -078]	-1.03	[-1.511, -551]	-1.29	[-1.796, -786]
Socialising	1.78	[1.041, 2.521]	3.42	[2.395, 4.436]	2.45	[1.632, 3.272]	2.21	[1.435, 2.994]	1.59	[902, 2.275]
Problem solving	1.59	[912, 2.26]	1.66	[1.026, 2.294]	1.24	[657, 1.824]	1.08	[514, 1.649]	0.78	[235, 1.321]
Writing	0.93	[418, 1.451]	-0.86	[-1.324, -404]	-1.43	[-1.94, -914]	-1.89	[-2.466, -1.317]	-2.18	[-2.8, -1.557]
Teach parts of the body	0.83	[372, 1.28]	0.08	[-325, 483]	-0.59	[-1.006, -173]	-0.88	[-1.313, -445]	-1.01	[-1.458, -57]
Teach about personal hygiene and body care	1.93	[1.129, 2.724]	3.03	[2.069, 3.983]	2.61	[1.736, 3.475]	2.28	[1.462, 3.087]	1.64	[919, 2.352]
Artistic expression	0.5	[132, 866]	1.12	[696, 1.545]	-0.08	[-452, 294]	-0.73	[-1.124, -332]	-0.94	[-1.346, -524]
Body language	0.96	[492, 1.417]	1.13	[659, 1.598]	0.01	[-402, 43]	-0.18	[-592, 242]	-0.37	[-788, 054]
Concentration	0.74	[331, 1.147]	0.6	[187, 1.005]	-0.26	[-655, 14]	-0.52	[-925, -115]	-0.72	[-1.131, -302]
Gross motor coordination	1.07	[602, 1.542]	1.45	[94, 1.967]	0.4	[-047, 84]	0.06	[-376, 498]	-0.16	[-592, 282]
Fine motor coordination	1.24	[737, 1.75]	1.3	[784, 1.817]	0.41	[-054, 873]	-0.24	[-697, 219]	-0.56	[-1.031, -096]
Gender identification	1.11	[599, 1.618]	0.43	[-012, 88]	-0.21	[-649, 229]	-0.38	[-819, 066]	-0.59	[-1.043, -143]
Responsibility	1.27	[721, 1.811]	1.76	[1.174, 2.349]	1.28	[747, 1.822]	1.09	[572, 1.615]	0.62	[129, 1.105]
Speech/story telling	1.45	[846, 2.046]	0.74	[236, 1.248]	0.24	[-244, 727]	-0.25	[-737, 232]	-0.44	[-925, 053]
Explore the community	0.4	[-118, 909]	-1.51	[-1.99, -1.035]	-3.03	[-3.875, -2.193]	-3.23	[-4.142, -2.313]	-3.46	[-4.477, -2.447]
Explore regional culture	0.5	[-007, 1.01]	-1.68	[-2.204, -1.162]	-2.67	[-3.397, -1.945]	-3.1	[-3.961, -2.243]	-3.52	[-4.551, -2.495]
Describe as being a main duty										
Prepare in assessment reports about children	0.12	[-378, 609]	1.53	[1.064, 1.991]						
Prepare in descriptive reports about classroom activities	-0.24	[-787, 314]	-1.85	[-2.372, -1.326]						
Support the development of goals at the centre	-0.12	[-71, 474]	-2.06	[-2.616, -1.497]						
Contribute to design and implementation of teaching strategies	-0.04	[-702, 626]	-2.32	[-2.941, -1.701]						
Perform work with families of the children	0.44	[008, 872]	-0.58	[-964, -192]						
Organise teaching materials	0.05	[-1.441, 1.548]	-4.1	[-5.505, -2.704]						

Notes: Table presents estimated parameters for IRT measurement model of TAs' learning activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2).

Table C.7: Measurement Model Parameters: TA Care Activities

Item (j)	β_j	α_{j1}	α_{j2}	α_{j3}	α_{j4}					
<i>Number of times did last week</i>										
Hygiene routines and care such as changing nappies, brush teeth, wash hands	1.11	[.631,1.578]	3.35	[2.443,4.256]	3.19	[2.331,4.057]	2.93	[2.131,3.721]	1.94	[1.337,2.549]
Supply medicines / remedies	1.11	[.631,1.578]	-3.91	[-5.029,-2.789]	-4.67	[-6.184,-3.154]				
Pamper children	1.11	[.631,1.578]	4.27	[3.01,5.534]	3.33	[2.431,4.237]	2.15	[1.511,2.788]		
Watch TV with children	1.11	[.631,1.578]	-3.32	[-4.228,-2.405]	-4.28	[-5.57,-2.982]	-5.45	[-7.539,-3.359]		
<i>Describe as being a main duty</i>										
Perform personal hygiene activities with the children	1.11	[.631,1.578]	2	[1.385,2.607]						
Clean the classroom	1.11	[.631,1.578]	-4.69	[-6.216,-3.171]						

Notes: Table presents estimated parameters for IRT measurement model of TAs' caring activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2). To aid convergence given the small number of items we restrict the discrimination parameter to be equal across items.

Table C.8: Measurement Model Parameters: Baseline Teacher Learning Activities

Item (j)	β_j	α_{j1}	α_{j2}	α_{j3}	α_{j4}					
Number of times did last week										
Read stories	0.887	[.712,1.061]	1.97	[1.751,2.188]	1.463	[1.271,1.655]	0.523	[.359,.688]	0.192	[.032,.352]
Tell stories	1.01	[.831,1.19]	1.341	[1.15,1.531]	0.754	[.581,.927]	0.053	[-.111,-.218]	-0.227	[-.392,-.062]
Conversation	2.982	[2.275,3.689]	6.992	[5.799,8.185]	6.659	[5.539,7.778]	6.224	[5.191,7.256]	4.392	[3.589,5.195]
Sing	2.246	[1.754,2.739]	6.527	[5.469,7.586]	5.879	[5.6,7.59]	5.041	[4.326,5.757]	3.564	[3.014,4.114]
Dance	0.758	[.593,.924]	1.77	[1.567,1.973]	1.369	[1.186,1.553]	0.537	[.378,.695]	0.163	[.009,.317]
Watch a video or educational programme on TV.	0.209	[.008,.41]	-1.701	[-1.892,-1.511]	-3.716	[-4.162,-3.271]	-4.423	[-5.048,-3.798]	-4.784	[-5.529,-4.039]
isit other places in the community	1.04	[.836,1.243]	2.113	[1.876,2.351]	1.743	[1.526,1.96]	1.196	[1.002,1.391]	0.825	[.642,1.008]
Free play within the nursery premises	0.722	[.565,.88]	0.98	[.814,1.146]	0.452	[.298,.607]	-0.045	[-.196,.107]	-0.352	[-.506,-.199]
Free play in the recreation area	0.942	[.755,1.129]	1.793	[1.581,2.005]	1.368	[1.175,1.56]	0.882	[.706,1.059]	0.49	[.322,.658]
Physical activities such as running, jumping	1.512	[1.241,1.783]	3.437	[3.061,3.814]	2.96	[2.627,3.293]	2.169	[1.889,2.449]	1.523	[1.275,1.77]
Group learning activities	1.218	[1.1,1.436]	1.905	[1.673,2.138]	1.649	[1.428,1.869]	1.133	[.932,1.334]	0.702	[.514,.891]
Teach colours	1.306	[1.093,1.52]	0.889	[.698,1.079]	0.485	[.301,.668]	0.117	[-.064,.297]	-0.19	[-.371,-.009]
Teach numbers	0.823	[.643,1.003]	-0.449	[-.609,-.29]	-0.647	[-.81,-.484]	-0.9	[-.107,-.73]	-1.099	[-.1,275,-.923]
Teach letter of the alphabet	0.784	[.58,.988]	-1.334	[-1.524,-1.145]	-1.571	[-1.772,-1.37]	-1.899	[-2.12,-1.679]	-2.043	[-2.273,-1.813]
Teach forms and shapes	1.398	[1.182,1.614]	0.96	[.763,1.156]	0.411	[.224,.598]	-0.114	[-.299,.071]	-0.468	[-.656,-.281]
Socialising	2.189	[1.765,2.613]	4.505	[3.933,5.076]	4.165	[3.626,4.703]	3.73	[3.231,4.23]	2.743	[2.32,3.166]
Problem solving	1.655	[1.366,1.944]	2.578	[2.267,2.889]	2.281	[1.986,2.576]	2.025	[1.742,2.307]	1.449	[1.194,1.704]
Writing	0.812	[.628,.995]	-0.829	[-.997,-.66]	-1.076	[1.252,-.899]	-1.557	[-1.754,-1.359]	-1.757	[-1.966,-1.549]
Teach parts of the body	1.368	[1.141,1.596]	1.83	[1.597,2.064]	1.418	[1.201,1.635]	0.912	[.71,1.114]	0.551	[.357,.745]
Teach about personal hygiene and body care	2.479	[2.002,2.955]	4.695	[4.08,5.31]	4.434	[3.844,5.025]	3.792	[3.255,4.329]	2.886	[2.418,3.354]
Artistic expression	1.265	[1.067,1.463]	1.484	[1.278,1.69]	0.657	[.475,.84]	-0.143	[-.32,.034]	-0.469	[-.649,-.289]
Body language	1.31	[1.093,1.527]	2.217	[1.967,2.467]	1.324	[1.116,1.532]	0.752	[.559,.944]	0.408	[.221,.594]
Concentration	1.808	[1.532,2.084]	1.682	[1.43,1.935]	1.244	[1.006,1.482]	0.798	[.572,1.024]	0.448	[.229,.667]
Gross motor coordination	1.334	[1.097,1.571]	2.586	[2.302,2.87]	2.185	[1.926,2.443]	1.471	[1.246,1.696]	1.054	[.844,1.264]
Fine motor coordination	1.41	[1.183,1.637]	2.304	[2.043,2.564]	1.927	[1.686,2.168]	0.994	[.788,1.199]	0.533	[.337,.728]
Gender identification	1.544	[1.294,1.794]	1.864	[1.617,2.112]	1.418	[1.187,1.648]	1.001	[.784,1.218]	0.673	[.465,.881]
Responsibility	2.059	[1.712,2.406]	2.552	[2.209,2.896]	2.367	[2.034,2.701]	2.109	[1.788,2.43]	1.575	[1.282,1.869]
Speech/story telling	1.697	[1.419,1.975]	2.467	[2.171,2.763]	2.051	[1.776,2.325]	1.579	[1.326,1.833]	1.075	[.841,1.31]
Explore the community	0.641	[.46,.821]	-1.028	[-1.196,-.859]	-1.969	[-2.186,-1.752]	-2.287	[-2.528,-2.045]	-2.562	[-2.829,-2.295]
Explore regional culture	0.841	[.643,1.038]	-1.209	[-1.394,-1.023]	-2.236	[-2.48,-1.991]	-2.711	[-2.997,-2.425]	-2.862	[-3.164,-2.561]

Notes: Table presents estimated parameters for IRT measurement model of teachers' learning activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2).

Table C.9: Measurement Model Parameters: Baseline Teacher Care Activities

Item (j)	β_j	α_{j1}	α_{j2}	α_{j3}	α_{j4}
Number of times did last week	1.21	[4.411, 5.807]	4.77	[4.149, 5.381]	4.39
Hygiene routines and care such as changing nappies, brush teeth, wash hands	1.21	[-2.77, -2.186]	-2.89	[-3.216, -2.565]	-3.13
Supply medicines / remedies	1.21	[4.16, 5.406]	4.58	[3.999, 5.161]	4.35
Pamper children	1.21	[-2.414, -1.879]	-3.68	[-4.096, -3.272]	-4.19
Watch TV with children	1.21	[.963, 1.464]	5.11	[.963, 1.464]	2.95
		[.963, 1.464]	-2.48	[-3.48, -2.784]	-3.33
		[.963, 1.464]	4.78	[3.813, 4.889]	2.88
		[.963, 1.464]	-2.15	[-4.679, -3.699]	-4.46

Notes: Table presents estimated parameters for IRT measurement model of teachers' caring activities at endline alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items in the "main duty" block are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. Items relating to the number of days on which the teacher performs difference activities are ordinal and we model them using the graded response model described in equation (4.2). To aid convergence given the small number of items we restrict the discrimination parameter to be equal across items.

Table C.10: Measurement Model Parameters: ECERS-R Direct Observations of Classroom Quality

Item (j)	β_j	α_j
Item 9, Sub-Item 1	1.02 [.53, 1.5]	2.35 [1.97, 2.84]
Item 9, Sub-Item 3	0.06 [-.29, .46]	-0.53 [-.92, -.2]
Item 10, Sub-Item 1	-1.27 [-4.9, -.67]	4.44 [3.68, 8.51]
Item 10, Sub-Item 3	0.87 [.51, 1.37]	0.11 [-.2, .39]
Item 10, Sub-Item 4	0.13 [-.55, .57]	2.34 [2.09, 2.82]
Item 11, Sub-Item 1	-0.78 [-2.4, -.18]	3.03 [2.55, 4.7]
Item 11, Sub-Item 2	0.79 [.44, 1.24]	-0.13 [-.52, .22]
Item 11, Sub-Item 3	-0.37 [-1.44, .21]	3.09 [2.71, 4]
Item 12, Sub-Item 1	1.08 [.66, 1.51]	1.18 [.76, 1.67]
Item 12, Sub-Item 2	0.76 [.43, 1.21]	-0.88 [-1.28, -.59]
Item 12, Sub-Item 3	0.86 [.56, 1.3]	-1.02 [-1.36, -.8]
Item 12, Sub-Item 4	0.16 [-.34, .52]	1.33 [1.08, 1.69]
Item 13, Sub-Item 1	1.19 [.73, 1.68]	1.99 [1.59, 2.44]
Item 14, Sub-Item 1	0.15 [-.15, .43]	1.14 [.9, 1.44]
Item 14, Sub-Item 2	2.24 [1.68, 3.26]	-0.50 [-1.34, .13]
Item 14, Sub-Item 3	0.49 [.01, .9]	1.74 [1.43, 2.16]
Item 15, Sub-Item 1	0.96 [.55, 1.46]	-0.13 [-.46, .14]
Item 15, Sub-Item 2	1.00 [-.48, 1.41]	3.86 [3.18, 4.16]
Item 16, Sub-Item 1	1.52 [1.13, 2.11]	0.39 [-.05, .84]
Item 16, Sub-Item 2	1.46 [1, 2.06]	2.69 [2.08, 3.51]
Item 17, Sub-Item 1	2.09 [1.43, 3.11]	3.71 [2.85, 5.18]
Item 17, Sub-Item 2	1.76 [1.17, 2.52]	1.75 [1.13, 2.5]
Item 18, Sub-Item 1	2.63 [1.7, 4.4]	5.00 [3.79, 7.45]
Item 18, Sub-Item 2	1.08 [.55, 1.8]	3.94 [3.27, 4.57]
Item 18, Sub-Item 3	1.27 [.9, 1.75]	-0.15 [-.54, .18]
Item 19, Sub-Item 1	0.34 [-.52, .99]	3.43 [3.03, 4.12]
Item 19, Sub-Item 2	1.07 [.7, 1.55]	1.62 [1.12, 2.32]
Item 20, Sub-Item 1	0.99 [.65, 1.52]	0.66 [.38, 1.01]
Item 20, Sub-Item 2	0.98 [.63, 1.46]	-0.13 [-.61, .3]
Item 22, Sub-Item 1	-0.21 [-.52, .08]	1.06 [.78, 1.41]
Item 23, Sub-Item 1	-0.36 [-.87, .09]	3.13 [2.7, 3.69]
Item 23, Sub-Item 2	1.26 [.8, 1.84]	2.03 [1.65, 2.63]
Item 24, Sub-Item 1	0.67 [.34, 1.07]	0.25 [-.02, .52]
Item 25, Sub-Item 1	2.41 [1.81, 3.94]	1.43 [.81, 2.33]
Item 26, Sub-Item 1	3.26 [2.14, 5.51]	-0.60 [-1.74, .38]
Item 26, Sub-Item 2	0.32 [.02, .71]	-1.10 [-1.47, -.82]
Item 27, Sub-Item 2	2.03 [1.23, 3.77]	0.67 [.06, 1.34]
Item 28, Sub-Item 1	1.06 [.63, 1.54]	1.82 [1.38, 2.42]
Item 28, Sub-Item 2	1.26 [.55, 1.84]	3.96 [2.97, 4.28]
Item 29, Sub-Item 1	1.63 [1.04, 2.43]	2.45 [1.81, 3.42]
Item 29, Sub-Item 2	1.51 [.9, 2.04]	3.82 [2.94, 4.26]
Item 30, Sub-Item 1	0.52 [.03, 1.02]	2.08 [1.75, 2.58]
Item 30, Sub-Item 2	1.96 [1.36, 3.28]	4.11 [3.28, 5.92]
Item 31, Sub-Item 2	2.50 [1.18, 3.36]	5.14 [3.21, 5.94]
Item 31, Sub-Item 3	1.32 [.82, 1.86]	2.78 [2.25, 3.44]
Item 32, Sub-Item 1	1.03 [.47, 1.53]	3.44 [2.85, 4.16]
Item 32, Sub-Item 3	2.05 [1.4, 2.86]	3.57 [2.78, 4.57]
Item 33, Sub-Item 1	0.22 [-.25, .55]	2.36 [2.09, 2.77]
Item 33, Sub-Item 2	0.81 [.37, 1.16]	2.53 [2.12, 3.1]
Item 33, Sub-Item 3	-0.18 [-.72, .15]	3.09 [2.66, 3.63]
Item 35, Sub-Item 1	21.41 [8.19, 33.44]	-7.23 [-12.91, -1.54]
Item 35, Sub-Item 2	0.65 [.33, .96]	0.69 [.37, 1.06]
Item 36, Sub-Item 1	0.36 [.06, .74]	0.70 [.48, .98]
Item 36, Sub-Item 2	1.84 [1.5, 2.56]	2.80 [2.19, 3.87]
Item 38, Sub-Item 1	0.88 [.55, 1.31]	1.32 [.94, 1.77]
Item 38, Sub-Item 2	4.54 [3.12, 9.14]	0.01 [-1.85, 1.36]
Item 39, Sub-Item 1	-0.74 [-1.32, -.37]	2.95 [2.38, 3.84]
Item 39, Sub-Item 2	0.87 [.28, 1.39]	3.75 [3.21, 4.34]
Item 40, Sub-Item 1	-0.06 [-.58, .37]	1.61 [1.34, 2.1]
Item 40, Sub-Item 2	1.02 [.53, 1.5]	2.35 [1.97, 2.84]
Item 40, Sub-Item 3	0.06 [-.29, .46]	-0.53 [-.92, -.2]

Notes: Table presents estimated parameters for IRT measurement model of the ECERS-R directly observed teaching quality alongside 95% confidence intervals in brackets. Parameters are estimated only using observations in the control group. All items are binary and we model them using the IRT model described in equation (4.1) with the guessing parameter set to 0. All items are reverse scores such that a 1 indicates better observed teaching processes and a 0 worse.

D Differential-Item Functioning

Whether the measures that researchers use to capture child development are invariant to the inputs of the child development production function (such as age, gender and material and time investments) is a crucial question in estimating and interpreting the impacts of inputs on child development (Heckman et al., 2013, 2020; Heckman and Zhou, 2021). If such inputs alter the mapping from underlying latent skills to response patterns, then not accounting for this will lead to bias in estimates of the causal effects of these inputs on latent skills.

In this appendix, we test for differential item functioning by treatment status for every item separately. As set out in Section 4.1, for each item j , we estimate parameters $\psi_j^C = [\alpha_j, \beta_j, g_j]$ on the control group. We consider it desirable to use only the control group to estimate the parameters of the measurement model *even* in the case of measurement invariance since this makes it straightforward to assume normality of the underlying latent factor in the control group but not make this assumption in the treatment groups.²⁷ Here, therefore, we are testing the null hypothesis that $H_0 : \psi_j^C = \psi_j^T$ for $T \in HIM, HIM+FE$. We do this under the maintained assumption that all other parameters are the same across treatment groups. Specifically, for every item j :

1. Let $\hat{\psi}_j^C = [\hat{\alpha}_j, \hat{\beta}_j, \hat{g}_j]$ denote our estimates of the parameters associated with item j estimated on the control group. Let the associated estimated variance-covariance matrix be $\hat{\Sigma}_j^C$.
2. We estimate the measurement model using only respondents in the HIM treatment group, imposing measurement invariance in all parameters other than those associated with item j and other than the mean and variance of the latent distribution.²⁸ Let the estimates of the item- j parameters in the HIM group be $\hat{\psi}_j^{HIM}$ and the associated variance-covariance matrix be $\hat{\Sigma}_j^{HIM}$.
3. We use our estimated difference in parameters $\hat{\psi}_j^C - \hat{\psi}_j^{HIM}$ and the associated variance-covariance matrix $\hat{\Sigma}_j^C + \hat{\Sigma}_j^{HIM}$ to test the null hypothesis $H_0 : \psi_j^C = \psi_j^{HIM}$ using a Wald test. We call the associated p-value p_j^{HIM} .

We repeat the procedure for the HIM+FE treatment group to obtain p_j^{HIM+FE} . We do this for both our measurement models of cognitive development and socio-emotional problems.

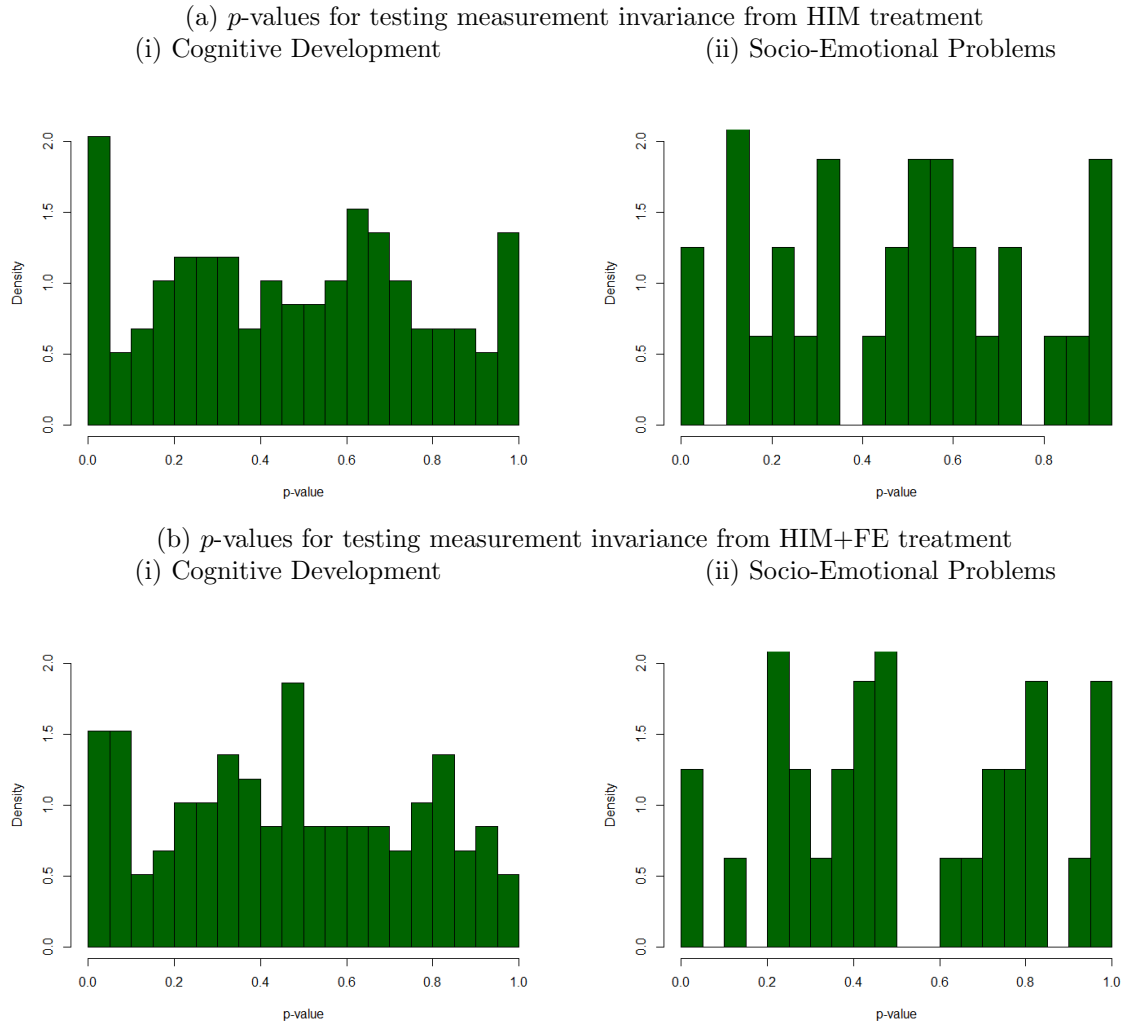
We plot the distribution of these p -values across all items in Figure D.1. For both treatment arms, and across both cognitive development and socio-emotional problems, the estimated p -values are roughly uniformly distributed in the unit interval. This is what we would expect if the true underlying parameters were invariant across groups. We take this as evidence that differential item functioning is not a major concern in our setting and, therefore, impose measurement invariance in estimating the factors that we use in our main results.

We additionally perform two robustness checks. First, we drop all items where the p -value obtained from testing the null of measurement invariance was less than 0.05 – for cognitive development, this amounts to

²⁷Many plausible patterns of heterogeneous impacts would rule out normality in the treatment groups.

²⁸Note that this also imposes normality of the latent factor in the treatment group which is an assumption that we do not need to make in estimating our main results.

Figure D.1: p -values for testing the null hypothesis of measurement invariance for all items separately



12 out of a total of 118 items in the HIM treatment group and 9 out of 118 in the HIM+FE treatment group – and re-estimate the factor model without these items. Second, we estimate a factor model that includes all items but allows parameters for this same set of items to be different in the treatment group in question. Table D.1 shows that our impacts remain very similar in both cases.

Table D.1: Robustness to Allowing for Differential Item Functioning in Estimating Impacts on Cognitive and Socioemotional Factor Scores

	(1)	(2)	(3)	(4)	(5)	(6)
	Cognitive development			Socioemotional problems		
HIM	-0.022 (0.088) p=0.786	-0.018 (0.087) p=0.817	-0.021 (0.087) p=0.800	-0.015 (0.150) p=0.930	-0.026 (0.160) p=0.865	-0.005 (0.156) p=0.969
HIM+FE	0.137* (0.073) p=0.053	0.145** (0.073) p=0.040	0.146** (0.074) p=0.040	-0.133 (0.169) p=0.437	-0.170 (0.177) p=0.345	-0.176 (0.179) p=0.329
Difference	0.159** (0.067) p=0.017	0.163*** (0.066) p=0.007	0.167*** (0.065) p=0.007	-0.118 (0.150) p=0.441	-0.143 (0.158) p=0.362	-0.172 (0.156) p=0.287
N	1074	1074	1074	1074	1073	1074
<i>Items removed:</i>						
$p_{HIM}^{DIF} \leq 0.05$	Y	N	N	Y	N	N
$p_{HIM+FE}^{DIF} \leq 0.05$	N	Y	N	N	Y	N
<i>Items with treatment-specific measurement parameters:</i>						
$p_{HIM}^{DIF} \leq 0.05$	N	N	Y	N	N	Y
$p_{HIM+FE}^{DIF} \leq 0.05$	N	N	Y	N	N	Y

Note. Single-hypothesis two-sided p -values calculated using a cluster bootstrap, resampling triplets with replacement (1,000 iterations). Standard errors (bootstrapped) 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. Columns (1) and (4) drop items where the estimated p -value for testing the null hypothesis of measurement invariance to the HIM treatment was less than 0.05. Columns (2) and (5) do the same for the HIM+FE treatment. Columns (3) and (6) includes all items but allow measurement parameters to vary for the items where our p -values for testing measurement invariance are less than 0.05.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E Modelling Appendix

In this appendix we provide a formal exposition of the model we refer to in Section 5. This model is flexible and incorporates a range of different mechanisms through which the HIM and HIM+FE programs might impact child development. As such, its purpose is to provide a framework which motivates the interpretation we propose for our results rather than to generate an unambiguous set of predictions about the impacts of these different programs.

We start by examining what happens when teachers are given additional resources, in the form of teaching assistants, in a context such as ours where teachers have a high degree of autonomy over what they do with

their time and work more than their contractual hours (see Section 2 for discussion of these features). We show that teachers may respond by changing the time that they devote to different types of activities in the pre-school and that the sign of these effects is, in general, ambiguous since the additional resources give rise to three distinct effects that may operate in opposite directions. We then incorporate the possibility that the FE intervention might lead to a shift in teachers' perceived production function. This shift may come about as a result of *real* changes in productivity with which teachers can accomplish certain tasks or through them updating their perceptions of the production function even if the underlying function itself remains constant. In particular, consider three types of changes that we think may have been caused by the FE curriculum (see Section 2). These include changes in teachers' perceptions about the optimal mix of learning and caring activities, the substitutability of teachers and TAs in performing different activities, and the overall productivity of what happens in the classroom for child development.

E.1 Teachers' Time-Use and the Process of Child Development

In line with the two categories of classroom activities we observe in our data, in our model teachers allocate their total time in the classroom, N , between learning ($L_t = \tau_t N$) and care ($C_t = (1 - \tau_t)N$) activities, where $\tau_t \in [0, 1]$ is the fraction of teachers' time spent on learning activities. Teachers' total endowment of time, which is normalised to 1, is divided between classroom time (N), which involves direct contact with children, and other time, denoted by K , which includes time spent on leisure and/or administrative tasks for the preschool. This gives us the constraint:

$$1 = K + N = K + L_t + C_t = K + \tau_t N + (1 - \tau_t)N \quad (\text{E.1})$$

We use H to denote child development and assume that it is a function of “learning” and “care” activities, L_t and C_t , performed by teachers and of the availability of teaching assistants, A :

$$H = f(L_t, C_t, A) \quad (\text{E.2})$$

We denote the partial derivative of f with respect to input $i \in \{1, 2, 3\}$ as f_i and the cross-partial derivative with respect to inputs $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3\}$ is f_{ij} . We assume that the f is continuous, increasing in all arguments, and concave. We further assume that $f_{12} > \max(f_{11}, f_{22})$, i.e. that any substitutability between teachers' learning and care activities is quantitatively smaller than the rate at which the marginal product of these inputs diminishes.²⁹ This condition is always satisfied if care and learning activities are q-complementary (i.e. if $f_{12} > 0$) and can hold even when they are substitutes.

We assume that teachers care about children's development H . However, their utility, $u(., .)$, also depends positively on the amount of time they do *not* spend in the classroom and thus can spend on leisure and/or administration, K .

²⁹It should be noted that formally we only require this locally around the chosen optimum. This condition holds under commonly-used production functions including Cobb-Douglas. Locally around the optimum, it also holds with a concave CES production function with equal factor prices (which holds in our case since the shadow cost of learning time and care time are equal).

$$u(H, K) = u(H, 1 - L_t - C_t) \quad (\text{E.3})$$

We denote the first- and second-order partial derivatives of u as u_i and u_{ij} respectively where $i \in \{H, K\}$ and $j \in \{H, K\}$. We assume $u(\cdot, \cdot)$ is continuous, increasing in both arguments and is concave. We further assume that teachers' preferences are separable in H and K , so $u_{HK} = 0$.

Teachers choose C_t and L_t taking A as given, to solve the following problem:

$$\max_{L_t, C_t} u(f(L_t, C_t, A), 1 - L_t - C_t) \quad (\text{E.4})$$

subject to the production function in equation (E.2). This gives two first-order conditions:

$$0 = u_H f_1 - u_K \quad (\text{E.5})$$

$$0 = u_H f_2 - u_K \quad (\text{E.6})$$

Combining these expressions, we have that at the chosen optimum:

$$f_1 = f_2 = \frac{u_K}{u_H}$$

E.2 Impacts of an Increase in Teaching Assistants' Time

We now explore the channels through which the HIM program – an exogenous increase in TA time – might affect teachers' time-use.

Let L_t^* and C_t^* denote, respectively, teachers' optimal choices for learning and care activities. By differentiating these FOCs (equations (E.5) and (E.6)) with respect to TA time, A , we can study how the optimal choices of L_t and C_t vary following an exogenous increase in TA time:

$$0 = u_{HH} f_1 \left(f_1 \frac{dL_t^*}{dA} + f_2 \frac{dC_t^*}{dA} + f_3 \right) + u_H \left(f_{11} \frac{dL_t^*}{dA} + f_{12} \frac{dC_t^*}{dA} + f_{13} \right) + u_{KK} \frac{dL_t^*}{dA} + u_{KK} \frac{dC_t^*}{dA} \quad (\text{E.7})$$

$$0 = u_{HH} f_2 \left(f_1 \frac{dL_t^*}{dA} + f_2 \frac{dC_t^*}{dA} + f_3 \right) + u_H \left(f_{12} \frac{dL_t^*}{dA} + f_{22} \frac{dC_t^*}{dA} + f_{23} \right) + u_{KK} \frac{dL_t^*}{dA} + u_{KK} \frac{dC_t^*}{dA} \quad (\text{E.8})$$

Combining these expressions and using the fact that at the optimum we have $f_1 = f_2 = u_K/(u_H)$, we get that:

$$\frac{dL_t^*}{dA} = \frac{f_{12} - f_{22}}{X} \left[\underbrace{u_{HH} \frac{u_K}{u_H} f_3}_{(1)} + \underbrace{u_H \left(f_{13} + f_{12} \frac{f_{23} - f_{13}}{f_{12} - f_{22}} \right)}_{(2)} + \underbrace{\frac{f_{13} - f_{23}}{f_{12} - f_{22}} \left(-u_{HH} \left(\frac{u_K}{u_H} \right)^2 - u_{KK} \right)}_{(3)} \right] \quad (\text{E.9})$$

$$\frac{dC_t^*}{dA} = \frac{f_{12} - f_{11}}{X} \left[\underbrace{u_{HH} \frac{u_K}{u_H} f_3}_{(1)} + \underbrace{u_H \left(f_{23} + f_{12} \frac{f_{13} - f_{23}}{f_{12} - f_{11}} \right)}_{(2)} + \underbrace{\frac{f_{23} - f_{13}}{f_{12} - f_{11}} \left(-u_{HH} \left(\frac{u_K}{u_H} \right)^2 - u_{KK} \right)}_{(3)} \right] \quad (\text{E.10})$$

where

$$X = - \left(u_{HH} \left(\frac{u_K}{u_H} \right)^2 + u_{KK} \right) (2f_{12} - f_{11} - f_{22}) + u_H (f_{11}f_{22} - (f_{12})^2) > 0$$

We consider first the overall effect of an increase in TA time on the amount of time teachers spend on learning activities. Equation (E.9) shows this overall effect is comprised of three effects, numbered (1) to (3):

1. A **resource effect**: This is always negative. It represents the fact that the new effort from TAs increases child development for the same level of effort from teachers, leading teachers to reallocate some of their time away from teaching activities into leisure.
2. A **complementarity/substitutability effect**. This describes how teachers shift their effort in learning activities due to the fact that the addition of TAs may change the marginal product of this effort. This effect is positive whenever the additional teaching assistant resources increase the marginal product of teachers' learning time. There is a direct component (i.e. f_{13} for learning) and an indirect component which comes from the fact that the TA resources may alter the amount of caring activities which, in turn, alters the marginal product of learning.³⁰
3. A **comparative advantage effect**: This effect means that teachers will reallocate their time to the activity (learning or care) that is more complementary with TAs time and away from the activity that is more substitutable. So if TAs' time is more substitutable with teachers' care time than learning time (i.e. $f_{13} > f_{23}$) which is what we might intuitively expect, then the comparative advantage effect will serve to increase teachers' effort in learning.

The overall impact of an increase in TA time on teachers' caring activities is comprised of three analogous effects set out in equation (E.10). The total amount of time teachers spend on classroom activities will only depend on the resource and complementarity effects:

$$\frac{d(L_t^* + C_t^*)}{dA} = \frac{1}{X} \left[u_{HH} \frac{u_K}{u_H} f_3 (2f_{12} - f_{11} - f_{22}) + u_H (f_{13}(f_{12} - f_{22}) + f_{23}(f_{12} - f_{11})) \right]$$

³⁰We are using complementarity and substitutability here to refer to two inputs i and j being q -complements (if $f_{ij} > 0$) or q -substitutes (if $f_{ij} < 0$) (Sato and Koizumi, 1973).

If teachers believe TA time to be substitutable with their own (in the sense that $f_{13} < 0$ and $f_{23} < 0$), an increase in TA time in the classroom will lead to an overall unambiguous reduction in the time that teachers spend with children. If teachers instead believe TA inputs and their own to be complementary (in the sense that $f_{13} > 0$ and/or $f_{23} > 0$), this effect will be ambiguous. In either case, they will reallocate their time in the classroom to the activity that they perceive to be most complementary with TA input.

E.3 Impacts of Teacher Training (FE)

In order to formalize the channels through which the addition of FE may have impacted teachers' time-use, we modify our model to explicitly allow for the possibility that training could change the *perceived* production function in the following ways: (a) an increase in total factor productivity; (b) a change in the relative productivity of learning vs. care activities; and (c) a change in the degree of substitutability between teachers and TAs in the performance of different types of activities in the classroom. Such changes may occur as a result of changes in the underlying production function if teachers are now able to perform various tasks more productively or because the training correcting misperceptions about various aspects of the production function.

We assume that teachers allocate TA time, A , between learning activities (L_a) and care activities (C_a):

$$A = L_a + C_a = \tau_a A + (1 - \tau_a)A \quad (\text{E.11})$$

where, as with teachers, τ_a represents the fraction of TA time that is devoted to learning activities. Modelling TA time in this way allows us to explicitly account for the possibility that the FE training program changes teachers' perceptions about which classroom activities they have a comparative advantage in relative to TAs.

Drawing on Caucutt et al. (2017), we work with a specific perceived production function for child development which is a special case of that considered above. In particular, we assume that teachers perceive that child development is produced by combining aggregates of learning and personal care activities. These aggregates are themselves determined by a CES aggregator of the time that teachers and TAs devote to each type of activity. In what follows, symbols with a \sim denote *perceived* parameters, emphasizing the fact that they *may* not be equal to the true parameters. Thus, the perceived process of child development is as follows:

$$H = \tilde{z}(\tilde{w}^{1-\rho}\tilde{L}^\rho + (1 - \tilde{w})^{1-\rho}\tilde{C}^\rho)^{\frac{1}{\rho}} \quad (\text{E.12})$$

where $0 < \tilde{z}$ is perceived total factor productivity and $0 < \tilde{w}$ represent the teachers' perceptions of the relative importance of learning as compared with care activities. ρ governs the elasticity of substitution. \tilde{C} and \tilde{L} are the aggregators of TA and teacher activities as perceived by the latter, determined by a CES aggregator:

$$\begin{aligned} \tilde{L} &= (\tilde{\theta}_l L_t^\lambda + (1 - \tilde{\theta}_l) L_a^\lambda)^{\frac{1}{\lambda}} = (\tilde{\theta}_l \tau_t^\lambda N^\lambda + (1 - \tilde{\theta}_l) \tau_a^\lambda A^\lambda)^{\frac{1}{\lambda}} \quad \lambda \in (0, 1] \\ \tilde{C} &= (\tilde{\theta}_c C_t^\lambda + (1 - \tilde{\theta}_c) C_a^\lambda)^{\frac{1}{\lambda}} = (\tilde{\theta}_c (1 - \tau_t)^\lambda N^\lambda + (1 - \tilde{\theta}_c) (1 - \tau_a)^\lambda A^\lambda)^{\frac{1}{\lambda}} \end{aligned} \quad (\text{E.13})$$

where $\tilde{\theta}_l$ and $\tilde{\theta}_c$ represent teachers' perceptions about the relative efficiency of teachers and TAs at performing learning and care activities, respectively. We note that the parameters of the aggregator functions in equation (E.13) are different for learning (\tilde{L}) and care (\tilde{C}) activities, allowing TAs to be better substitutes for teachers in one type of activities than the other. In particular, one might expect TAs to be better substitutes in care activities, which require less knowledge and training in early learning provision.³¹

In this specification of the model, teachers choose three variables, L_t , C_t and L_a ,³² or, equivalently, N , τ_t and τ_a , to maximise their objective function, subject to the perceived production function and taking A as given. Formulating the expression in terms of N , τ_t and τ_a is simpler and more intuitive, as, given that teachers' dis-utility from classroom time does not depend on whether they engage in learning or caring activities during this time, it allows us to consider a two-stage problem. In the two-stage problem, given N , teachers optimize the allocation of time to different activities. They decide how to determine N , given the utility and production functions.

Given the teacher's total hours spent in classroom activities, N , the second-stage problem gives rise to the following first-order conditions:

$$\tau_t : 0 = \frac{\tilde{z}}{\rho} h^{\frac{1}{\rho}-1} \left(\tilde{w}^{1-\rho} \frac{\partial \tilde{L}^\rho}{\partial \tau_t} + (1-\tilde{w})^{1-\rho} \frac{\partial \tilde{C}^\rho}{\partial \tau_t} \right) \quad (\text{E.14})$$

$$\tau_a : 0 = \frac{\tilde{z}}{\rho} h^{\frac{1}{\rho}-1} \left(\tilde{w}^{1-\rho} \frac{\partial \tilde{L}^\rho}{\partial \tau_a} + (1-\tilde{w})^{1-\rho} \frac{\partial \tilde{C}^\rho}{\partial \tau_a} \right) \quad (\text{E.15})$$

where $h = (\tilde{w}^{1-\rho} \tilde{L}^\rho + (1-\tilde{w})^{1-\rho} \tilde{C}^\rho)$. This last term, as well as \tilde{z} , cancels out from both first-order conditions. This implies that the ratios of both TA and teacher time are pinned down independently of \tilde{z} . i.e. \tilde{z} might affect *total* teacher time but will never change the ratio of learning to caring activities. Substituting in equation (E.14) the expressions for $\partial \tilde{L}^\rho / \partial \tau_t$ and $\partial \tilde{C}^\rho / \partial \tau_t$, we get:

$$\left(\frac{\tilde{w}}{1-\tilde{w}} \right)^{1-\rho} \frac{\tilde{\theta}_l}{\tilde{\theta}_c} = \left(\frac{h_c}{h_l} \right)^{\frac{\rho}{\lambda}-1} \left(\frac{\tau_t}{1-\tau_t} \right)^{1-\lambda} \quad (\text{E.16})$$

where $h_l = \tilde{\theta}_l \tau_t^\lambda N^\lambda + (1-\tilde{\theta}_l) \tau_a^\lambda A^\lambda$ and $h_c = \tilde{\theta}_c (1-\tau_t)^\lambda N^\lambda + (1-\tilde{\theta}_c) (1-\tau_a)^\lambda A^\lambda$. Analogously, considering equation (E.15), we obtain:

$$\left(\frac{\tilde{w}}{1-\tilde{w}} \right)^{1-\rho} \frac{1-\tilde{\theta}_l}{1-\tilde{\theta}_c} = \left(\frac{h_c}{h_l} \right)^{\frac{\rho}{\lambda}-1} \left(\frac{\tau_a}{1-\tau_a} \right)^{1-\lambda} \quad (\text{E.17})$$

Taking the ratio of equations (E.17) and (E.16), we obtain:

$$\frac{\tilde{\theta}_l}{1-\tilde{\theta}_l} \frac{1-\tilde{\theta}_c}{\tilde{\theta}_c} = \left(\frac{\tau_t}{\tau_a} \right)^{1-\lambda} \left(\frac{1-\tau_a}{1-\tau_t} \right)^{1-\lambda} \quad (\text{E.18})$$

Taking logs of expression (E.18), holding fixed N and totally differentiating with respect to \tilde{w} , gives the result that any change in teachers' perception of the relative importance of learning vs. care routines will,

³¹The restrictions on λ , which preclude the possibility that TAs and teachers are q -complements within either aggregator, guarantee that teaching assistants are not necessary for the production of child development.

³²As A is given exogenously, a choice of L_a determines C_a .

holding fixed their total time input, lead to a proportional change in both the fraction of time they and their TA allocate to learning. Both changes will always be of the same sign:

$$\frac{d\tau_a}{dw} = \frac{\tau_a(1 - \tau_a)}{\tau_t(1 - \tau_t)} \frac{d\tau_t}{d\tilde{w}} \quad (\text{E.19})$$

Taking logs of equation (E.16), holding N fixed and totally differentiating gives with respect to \tilde{w} gives:

$$\frac{1 - \rho}{\tilde{w}(1 - \tilde{w})} = \frac{\lambda - \rho}{\lambda} \frac{1}{h_l} \frac{dh_l}{d\tilde{w}} - \frac{\lambda - \rho}{\lambda} \frac{1}{h_c} \frac{dh_c}{d\tilde{w}} + \frac{1 - \lambda}{\tau_t(1 - \tau_t)} \frac{d\tau_t}{d\tilde{w}} \quad (\text{E.20})$$

which, after substituting in the expressions for $\frac{dh_l}{d\tilde{w}}$ and $\frac{dh_c}{d\tilde{w}}$ gives us that $\frac{d\tau_t}{d\tilde{w}} > 0$.³³ And, by expression (E.19), we have $\frac{d\tau_c}{d\tilde{w}} > 0$.

Discussion of FE channels We can now draw some implications for how FE may have impacted teacher behavior and map these to the effects we presented in Table 6. To recap, in that table we see the following changes in teacher and TA behavior in response to the addition of FE training to the HIM program: (a) increase in teacher overtime; (b) increase in the frequency with which teachers undertake learning activities; (c) no change in the frequency with which teachers undertake care activities; (d) no change in what TAs do in relation to either learning or care activities. We consider what mechanisms might be at play in generating this pattern of results by tracing out what we learn about the effects of changing \tilde{z} , $\tilde{\theta}_l$, $\tilde{\theta}_c$ and \tilde{w} on teacher and TA behavior through our model.

1. **Changing \tilde{z} :** As noted above, \tilde{z} cancels out from both first-order conditions in equations (E.14) and (E.15). This implies that *the optimal fractions of time allocated to different types of activities and between TAs and teachers are independent of the level of \tilde{z}* . Since we observe a change in teachers' relative time allocated to learning vs. caring activities in response to FE, the program must have had impacts beyond an increase in \tilde{z} , either real or perceived, alone.
2. **Changing \tilde{w} :** As we have seen, an increase in \tilde{w} – that is, in teachers' perceptions of the relative importance of learning activities as compared with care activities – will induce, conditional on a value of N , a proportional increase in both τ_t and τ_a , the fraction of time that both teachers and TAs devote to learning activities. Our empirical results indicate a change in teacher behavior, and a considerable increase in learning activities in HIM+FE preschools, relative to those with only HIM. This effect is consistent with an increase in \tilde{w} . However, we see no change in TA behavior.

33

$$\begin{aligned} \frac{dh_l}{d\tilde{w}} &= \tilde{\theta}_l N^\lambda \lambda \tau_t^{\lambda-1} \frac{d\tau_t}{d\tilde{w}} + (1 - \tilde{\theta}_l) A^\lambda \lambda \tau_a^{\lambda-1} \frac{d\tau_a}{d\tilde{w}} \\ &= \frac{d\tau_t}{d\tilde{w}} \left[\tilde{\theta}_l N^\lambda \lambda \tau_t^{\lambda-1} + (1 - \tilde{\theta}_l) A^\lambda \lambda \tau_a^{\lambda-1} \frac{\tau_a(1 - \tau_a)}{\tau_t(1 - \tau_t)} \right] \\ \frac{dh_c}{d\tilde{w}} &= -\tilde{\theta}_c N^\lambda \lambda (1 - \tau_t)^{\lambda-1} \frac{d\tau_t}{d\tilde{w}} - (1 - \tilde{\theta}_c) A^\lambda \lambda (1 - \tau_a)^{\lambda-1} \frac{d\tau_a}{d\tilde{w}} \\ &= -\frac{d\tau_t}{d\tilde{w}} \left[\tilde{\theta}_c N^\lambda \lambda (1 - \tau_t)^{\lambda-1} - (1 - \tilde{\theta}_c) A^\lambda \lambda (1 - \tau_a)^{\lambda-1} \frac{\tau_a(1 - \tau_a)}{\tau_t(1 - \tau_t)} \right] \end{aligned}$$

3. **Changing $\tilde{\theta}_l$ and/or $\tilde{\theta}_c$:** From equation (E.18), it is clear that a change in $\tilde{\theta}_l$ and/or $\tilde{\theta}_c$ would lead to a change in the relative allocation of teacher and TA time. One of the key foci of the FE curriculum was on the importance of high-quality learning time for children’s development, which is likely to map to an increase in $\tilde{\theta}_l$ in particular. The model implies that *an exogenous increase in teachers’ perceived comparative advantage relative to TAs in learning activities, i.e. in $\frac{\tilde{\theta}_l}{1-\tilde{\theta}_l} \frac{1-\tilde{\theta}_c}{\tilde{\theta}_c}$, will result in teachers spending a greater share of their time on learning activities relative to TAs.* This is consistent with what we see in the data: teachers increase the frequency with which they perform learning activities. However, an increase in $\tilde{\theta}_l$ relative to $\tilde{\theta}_c$ alone would also imply that TAs would increase the proportion of their time allocated to care activities (their comparative advantage) which we do not see.

Bringing these together suggests several insights relevant for interpretation of our findings. First, an increase in total factor productivity or perceived total factor productivity alone could not generate these changes. For this model to explain our results, it must be the case that FE increases teachers’ perceptions of the marginal product of their learning activities relative to their caring activities. This could be driven either through an increase in \tilde{w} , teachers’ perception of the overall usefulness of learning vs. care activities, or by an increase in $\tilde{\theta}_l$ relative to $\tilde{\theta}_c$, teachers’ perceived comparative advantage in learning activities. Formally, these two channels have opposite implications for the impacts on TA time use. An increase in the perceived overall usefulness of learning activities should lead teachers to also instruct their TAs to spend a higher proportion of their time on learning activities. However, an increasing perception of comparative advantage would suggest that teachers would direct TAs to increasingly exploit *their* comparative advantage in care activities. If both effects were at play, they could offset each other in their impacts on TA time use. Empirically, this is consistent with our observation that FE had no statistically significant impact on TA time use. However, these estimates are noisy and therefore we don’t draw any strong conclusions on the basis of them.