NBER WORKING PAPER SERIES

PRESCHOOL QUALITY AND CHILD DEVELOPMENT

Alison Andrew
Orazio Attanasio
Raquel Bernal
Lina Cardona Sosa
Sonya Krutikova
Marta Rubio-Codina

Working Paper 26191
http://www.nber.org/papers/w26191

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019, Revised May 2022

We thank Diana Pérez-Lopéz 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. 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). Ms Andrew and Dr Krutikova acknowledge funding from the ESRC Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. Prof. Bernal acknowledges funding from the British Academy Visiting Fellowship 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w26191.ack

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Alison Andrew, Orazio Attanasio, Raquel Bernal, Lina Cardona Sosa, Sonya Krutikova, and Marta Rubio-Codina. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Preschool Quality and Child Development
Alison Andrew, Orazio Attanasio, Raquel Bernal, Lina Cardona Sosa, Sonya Krutikova, and Marta Rubio-Codina
NBER Working Paper No. 26191
August 2019, Revised May 2022
JEL No. H43,I10,I20,J13

ABSTRACT

Global access to preschool has increased dramatically, yet preschool quality is often poor and evidence on how to improve it is scarce. We worked with the government of Colombia to implement a largescale randomized controlled trial evaluating two interventions targeting the quality of public preschools in Colombia. The first, which was designed by the government and rolled out nationwide, provided preschools with significant extra funding, mainly earmarked for hiring teaching assistants (TAs). The second additionally offered professional development training for existing teachers, delivered using a novel low-cost video-conferencing approach. We find that, despite increasing per-child expenditure by around a third, the first intervention did not improve child development and led to a reduction in the time that teachers spent in the classroom, including on learning activities. In contrast, the second intervention led to significant improvements in children's cognitive development, especially those from more disadvantaged backgrounds, at little extra cost. The addition of the professional development training offset the adverse effects of TA provision on the time teachers spent on learning activities in the classroom and improved the quality of teaching. When we interpret our results through the lens of a model of teacher behavior, two insights arise. First, income effects and a perception that TA time was a good substitute for their own may have led teachers to endogenously scale back their efforts in the classroom in response to the provision of new resources. Second, the training prompted teachers to increase their perception of the usefulness of learning activities for child development and their perception that they had a comparative advantage in these learning activities relative to the TAs.

Alison Andrew
Institute for Fiscal Studies
alison_a@ifs.org.uk

Orazio Attanasio
Department of Economics
Yale University
37 Hillhouse Avenue
New Haven, CT 06511
and Institute for Fiscal Studies, FAIR, BREAD and CEPR
and also NBER
orazio.attanasio@yale.edu

Raquel Bernal
Universidad de los Andes
and CEDE
rbernal@uniandes.edu.co

Lina Cardona Sosa
World Bank
1818 H St NW
Washington, DC 20433
United States
lcardonasosa@worldbank.org

Sonya Krutikova
Institute for Fiscal Studies
7 Ridgmount Street
London WC1E 7AE
United Kingdom
sonya_k@ifs.org.uk

Marta Rubio-Codin
Inter-American Development Bank
Social Protection and Health Division
Washington, DC 20577
martarubio@iadb.org

A controlled trials registry entry is available at https://www.socialscienceregistry.org/trials/1246

1 Introduction

It is now widely accepted that well-designed early child education (ECE) programs can have substantial and long-lasting positive effects on children (Elango et al., 2015). Consequently, there is significant momentum behind investing in early years education in both lower- and higher-income countries: universal access to quality early childhood care by 2030 is one of the Sustainable Development Goals and, globally, enrollment in pre-primary education is rising fast; it increased from 29% in 1990 to 49% in 2015.[1] However, as governments expand coverage of ECE programs, quality should be a first-order concern. If not of good quality, ECE programs may deliver few benefits for child development and can even be inferior to homecare (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). 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).[2] There is a need, therefore, 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 focus on understanding which aspects of ECE programs are most important for child development or on the effectiveness of specific improvements to existing programs. 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 and on potential mediating factors such as how teachers spend their time. We then set out a model of teacher behavior which helps us disentangle various mechanisms that may have generated the impacts that we find.

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 for the 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

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

[2] For example, many LMICs are resorting to adding pre-primary classes to existing primary schools without allocating sufficient extra resources or expertise to ensure that these are providing high-quality care and education tailored to the needs of young children (Neuman and Okeng'o, 2019).

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 0.17 of the control group standard deviation (SD); relative to the HIM-only arm, the addition of FE improved child development by 0.15 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 a third 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. The average teacher worked more than their contracted hours at baseline and had significant administrative duties. Therefore it is plausible that teachers might respond to the interventions by adjusting their total classroom time as well as adjusting how they split their classroom time between different teaching activities. Using novel data which capture teachers' day-to-day activities, we show that teachers responded to the HIM program by reducing the total amount of time that they allocated to their job. They reduced their involvement not only in care but also learning-focused activities which are highly correlated with children's development. The addition of FE, however, induced teachers to increase the time that they allocated to the job, increase their involvement in learning activities and improved the quality of teaching as directly observed by trained psychologists.

In order to interpret these findings, we set up a theoretical framework that allows us to consider how these two programs may have affected the preschool learning environment. In our model, teachers value child development and leisure. Child development is produced by combining learning and care-focused activities. Activities may be led either by a teacher or a TA although their productivities may differ. Teachers are free to allocate their own time and their TA's time (if they have one) as they see fit, given their preferences and their beliefs about the process of child development. Building on the approach in Caucutt, Lochner, and Park (2017), we explicitly allow teachers to hold incorrect beliefs about key parameters of the process of child development: total factor productivity, the relative importance of different activities for child development and the substitutability of TAs' and teachers' time. This approach allows us to analyze how the time-use of teachers and TAs might respond to teachers revising their beliefs about the process of child development.

Our model suggests that teachers' response to extra help in the classroom from TAs is a-priori ambiguous. There will be a negative resource effect whereby, holding teachers' effort constant, the addition of TAs increases child development and teachers react to this increase in the value of their endowment by increasing their leisure (or other activities) at the expense of their effort in the classroom. Further, the addition of TAs may either increase or decrease the (perceived) marginal product of teachers' effort depending on whether teachers and TAs are substitutes or complements in the perceived production function; this substitutability/complementarity effect may be either positive or negative. Finally, there is a comparative advantage effect whereby teachers will reallocate effort towards activities that they perceive to be more complementary with TAs' time. Thus, overall, teachers' response will be guided by how beneficial they *think* different

activities are for child development, as well as by their perception of how substitutable the TA input is with their own. An intervention that increases resources by providing TAs is most likely to be effective at improving quality of care if teachers believe that their own time is highly complementary with support from TAs and if teachers' marginal valuation of child development (relative to their marginal valuation of additional leisure time) is high. Our finding that teachers *reduced* their efforts in response to the additional TA support suggests that any perceived complementarities were not strong enough to overcome the negative resource effect in this context.

Our model highlights how interventions that change teachers' perceptions of the process of child development will alter teachers' own time-allocation, as well as how they utilise the time of their TAs. Interpreting our empirical evidence on time-use responses to the FE teacher training program through the lens of this model suggests that the training shifted teachers' beliefs in two ways: (i) it increased teachers' perception of the productivity of learning activities relative to care activities; and (ii) it strengthened teachers' perception of their comparative advantage in learning activities and TAs comparative advantage in care activities.

Taken together, the impacts on child development and teachers time allocation, suggest that, given teachers' preferences for different activities and outcomes 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 utilise 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 rigorously evaluate 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 ECE services rather than how to start them up.

Our paper contributes to several, more specific strands of the literature. The first looks at whether and how providing schools and preschools with additional resources improves the quality of the education they deliver (see Glewwe et al. (2011) and Evans and Popova (2016) for reviews). In particular, we examine a common approach to increasing resources: providing preschools and primary schools with teaching assistants. 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 was possibly 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 this crucial role of the training of auxiliary educational professionals for the precise role they are expected to 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, educational benefits followed. Our analysis suggests that the rollout of Colombia's nationwide quality improvement program led to similarly disappointing results as the Tennessee STAR experience. We offer a theoretical framework that formalizes when and how the provision of TAs can backfire if the TAs do not have clearly defined tasks and if teachers have scope to endogenously react to the increase in TAs by reducing their own effort. We provide empirical evidence that these mechanisms are important in 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). Yoshikawa et al. (2015) 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 the critical ingredients of effective preschool teacher training programs are.

Third, this paper speaks to a broader literature exploring how the investment decisions of key actors in the process of child development are shaped by their perception or beliefs about the production of child development, how misperceptions can lead to a sub-optimal level and mix of investments, and how targeted interventions can improve child development by correcting these misperceptions. While most of this literature has focused on parents' beliefs, we consider how teachers' beliefs shape their teaching practices. One strand of this literature has shown how misperceptions about children's current level of development and children's own effort in learning are common and these can distort parents' investment decisions (Dizon-Ross, 2019; Barrera-Osorio et al., 2020; Bergman, 2021; Kinsler and Pavan, 2021). Our paper speaks most directly to another strand on misperceptions about the production technology of child development itself. We build on work exploring how perceptions (and misperceptions) over the relative importance of different types of

inputs shape investment choices (Cunha, Elo, and Culhane, 2013; Caucutt et al., 2017; Boneva and Rauh, 2018; Attanasio, Cunha, and Jervis, 2019; Attanasio, Boneva, and Rauh, 2020a; Cunha, Elo, and Culhane, 2020). We then introduce a new type of misperception that we argue may be particularly important in our context - that over the comparative advantage of the different actors (in our case, teachers and TAs) in the child development process which can lead to gains from specialization not being fully exploited. Our experimental results add to the growing evidence that interventions targeting misperceptions over the child development process can be effective in both changing investment patterns and improving child development (Carneiro et al., 2021).

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. The estimates of the main impacts are presented in Section 5, alongside impacts on potential mediators. In Section 6, we set out a conceptual framework which helps us consider potential mechanisms more formally. 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.[3] 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 single largest pot of money was earmarked for hiring teaching assistants (TAs) to support the teachers. Prior

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

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.[4] 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 year later. Additionally, there was interest from a well-established Colombian NGO, *Fundación Éxito* (FE) and the Colombian National University, in offering a teacher training program in addition to 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 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 best-practice 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 and learning during class time and best practice in these.

The program was delivered through three components: (i) instruction through 16 monthly 3-hour-long sessions delivered via videoconferencing; (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 an important feature which enhances 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, required internet access 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.[5]

The HIM program cost the government a substantial amount: a 30% increase in per-child expenditure 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

[4]This 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.

[5]We 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.

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, which also had the largest number of HIs.[6] Randomization was at the level of the HI, with 40 HIs randomized into each of the three arms: (i) HIM, where preschools received the government quality improvement program; (ii) HIM+FE, where preschools received the teacher 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.[7]

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, from whom we drew a baseline sample of 15 to 17 children per HI.[8] Baseline data were collected between March and May 2013.[9] 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. As discussed in Section 4.2, some of the child development assessments (our key outcome measures) were unsuitable for children below the age of 48 months. Therefore, our main analysis sample comprises only children above 48 months at endline who were thus eligible for all assessments.

[6] The cities included are Bogotá, Cali, Medellín, Barranquilla, Bello, Palmira, Itagüí and Soledad.

[7] Our key hypotheses are set out in a pre-analysis plan held at the AEA trial registry (AEARCTR-0001246).

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

[9] HIs assigned to HIM and HIM+FE had already begun to make preparations for the HIM upgrades at the point of baseline. However, we do not see any imbalances that might be evidence of the program already having effects on child development.

3.1 Balance and Attrition

Attrition was relatively rare. We completed *some* endline child development assessments for all but 155 children (7.8%) of the 1,987 children in the baseline sample. Attrition was not related to treatment assignment (Table B.1). As discussed above and again in Section 4.2, 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.[10] This leaves us with 1075 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%. Likewise, attrition amongst this older group was not related to treatment assignment (Table B.1).

Table 1 shows baseline characteristics of our 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.[11]

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.

3.2 Compliance

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. However, 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.[12] 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 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 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

[10] In robustness Section 5.1.3 we show that the same patterns of our results hold when examining an "extended sample" that includes younger children for whom we have incomplete assessment data.

[11] When examining the coefficients on these control variables (in Table B.6), it is reassuring (given these slight imbalances) to note that baseline motor skills seem unimportant in predicting endline cognitive and socioemotional development. In contrast, baseline measures of language, problem solving and communication skills are highly predictive of endline outcomes.

[12] As detailed in Section 2, 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.

Table 1: Baseline Sociodemographic Characteristics and Child Development by Randomization Status for Analysis Sample

	Control	HIM	HIM+FE	HIM vs. Control *p*-value	HIM+FE vs. Control *p*-value	HIM vs. HIM+FE *p*-value	N
Male	0.456 (0.499)	0.552 (0.498)	0.524 (0.500)	*[p=0.004]*	*[p=0.085]*	*[p=0.493]*	1075
Age (months)	32.98 (2.120)	32.77 (2.179)	32.70 (2.279)	*[p=0.229]*	*[p=0.101]*	*[p=0.674]*	1075
HH income (million COP)	1333.1 (774.177)	1341.5 (777.608)	1338.3 (794.453)	*[p=0.923]*	*[p=0.966]*	*[p=0.872]*	1075
Mother's education (years)	12.63 (2.776)	12.37 (2.601)	12.67 (2.577)	*[p=0.302]*	*[p=0.923]*	*[p=0.201]*	1065
Father's education (years)	12.01 (3.041)	11.98 (3.116)	12.13 (3.068)	*[p=0.915]*	*[p=0.706]*	*[p=0.570]*	1004
Household size	3.385 (1.697)	3.477 (1.629)	3.213 (1.541)	*[p=0.501]*	*[p=0.172]*	*[p=0.078]*	1075
ASQ Communication	63.95 (19.765)	65.86 (20.842)	64.41 (20.150)	*[p=0.287]*	*[p=0.912]*	*[p=0.335]*	1075
ASQ Gross Motor	62.22 (21.669)	66.22 (20.886)	64.50 (20.005)	*[p=0.066]*	*[p=0.223]*	*[p=0.412]*	1075
ASQ Problem Solving	57.63 (19.507)	59.40 (20.347)	58.67 (19.186)	*[p=0.377]*	*[p=0.761]*	*[p=0.527]*	1075
ASQ Personal Social	57.86 (18.587)	60.49 (18.590)	58.98 (18.346)	*[p=0.154]*	*[p=0.523]*	*[p=0.302]*	1075
ASQ Fine Motor	46.98 (20.089)	51.50 (20.813)	46.67 (19.812)	*[p=0.073]*	*[p=0.807]*	*[p=0.016]*	1075
MacArthur-Bates Language	66.16 (24.088)	67.68 (24.025)	66.48 (23.377)	*[p=0.580]*	*[p=0.938]*	*[p=0.524]*	1075
ASQ Socio-Emotional	56.09 (21.420)	53.29 (19.734)	54.62 (20.617)	*[p=0.151]*	*[p=0.516]*	*[p=0.370]*	1075

Note. Baseline characteristics by treatment status for children included in the analysis sample (all children with complete child development assessment data at endline). Single-hypothesis two-sided *p*-values calculated using a block 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.

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 teach 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 training took place between June 2013 and June 2014. We have very limited data on implementation of this component. 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):

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

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, which we allow to be clustered at the level of the sampling triplet.

Pre-specified baseline controls for child-level outcomes include the child's age, age squared, gender, a set of city dummies and 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.[13]

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 single-hypothesis p-values using a block bootstrap, resampling the 40 randomization triplets

[13] We use the center average to ensure we have control variables defined 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 only including teachers who were present in the HI at baseline.

with replacement (1000 iterations).

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 (2016) which, building on Romano and Wolf (2005) and Romano and Wolf (2010), provides balanced asymptotic control of the family-wise error rate. In running the procedure, we use the block bootstrap described above, studentizing by the 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.

While we present impacts estimates on measures scored using the standard algorithms recommended by the test publishers, we also follow the literature in using structural measurement models to summarize the information contained in our measures efficiently; the advantages of the latter approach are discussed elsewhere, for instance, by Heckman et al. (2013). While these methods are not novel, it is useful to provide details on the specific approach we take. Therefore, we begin this section by outlining the specific 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 potential 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.

4.1 Measurement Model

We report impacts on outcome measures scored using the algorithms provided by the test developers. In addition, we adopt the increasingly common approach (see, e.g. Cunha, Heckman, and Schennach (2010); Heckman et al. (2013); Attanasio et al. (2020b); Agostinelli et al. (2021)) of using a structural measurement model to construct estimates of underlying latent factors capturing each outcome. These techniques combine the information contained in the available measures efficiently. Furthermore, when treatment effects are scaled relative to the variance of the control group, modelling measurement error directly allows for the

estimation of treatment effects that are unbiased since they allow the researcher to scale effects relative to true variation in the underlying construct uncontaminated by variability induced by measurement error.[14]

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. These methods – which have a long history in psychometrics (Van Der Linden and Hambleton, 1997) and are increasingly being used by economists (e.g. Das and Zajonc, 2010; Singh, 2020) – use non-linear linking functions (such as logit and ordered-logit models) to map indicators of responses to discrete items onto unobserved latent factors. In this sense, our specific model differs from linear factor models, which model multiple aggregated test scores as depending linearly on an underlying unobserved factor (e.g. Cunha et al., 2010; Heckman et al., 2013; Attanasio et al., 2020b; Agostinelli et al., 2021), but the underlying concepts are the same. Estimating underlying factors directly from the individual binary or ordinal item responses will yield efficiency gains (compared to a linear framework) if items vary substantially in their difficulty and discrimination power or if the official scoring algorithms were developed using samples from a population different from the one in which they are implemented. Both issues are relevant in our context.[15]

Specifically, let θ_{id} represent i's factor of interest in domain d where d can be cognitive development, socioemotional development, learning activities, care activities, or directly-observed classroom quality and where i represents either the individual child, teacher, TA or classroom. We assume that θ_{id} is normally distributed with zero mean and unit variance in the control group.[16] θ_{id}, however, is not observed directly. Instead, available measures, y_{ijd}, are noisy measures of the latent factor θ_{id}. Our measurement equations describe the ways that these latent factors determine the item responses. We estimate a dedicated measurement system for each domain, in that we assume that each item loads only one factor.

Depending on the nature of each item, we use one of three different specifications to map the underlying factor to item responses. 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 multiple-choice question. We model these items using a three-parameter "guessing" specification (Birnbaum, 1968) to describe the probability that i correctly answers item j:

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

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

[14] To see this, consider that child development is measured with error $\hat{\theta}_i = \theta_i + u_i$, where θ_i is the true underlying level of development, $\hat{\theta}_i$ is the observed measure, $E(u_i) = 0$ and $V(u_i) > 0$. Conceptually, our treatment effect of interest, scaled relative to the variance of underlying child development in the control group, is $\gamma \equiv \frac{E(\theta_i|T) - E(\theta_i|C)}{V(\theta_i|C)}$. A naive estimator is the difference in the observed measure of child development across treatment and control, scaled by the sample variance of the observed measure in the control group. However, under measurement error, this will be biased towards zero: $E(\hat{\gamma}) = \frac{E(\theta_i|T) - E(\theta_i|C)}{V(\theta_i|C) + V(u_i|C)} \neq \gamma$. A structural measurement model allows for the direct estimation of $V(\theta_i|C)$.

[15] For 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 (Schrank et al., 2005). This norming sample is likely to differ substantially from the children in our sample.

[16] In Appendix Table B.8, we show our results are robust to relaxing this normality assumption.

changes with the underlying factor. g_{jd} is the "pseudo-guessing parameter" and is the asymptotic probability of i choosing correctly as $\theta_{jd} \to -\infty$.

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.

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:

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

We estimate the measurement models by maximum likelihood using an Expectation-Maximization (EM) algorithm.[17] We estimate the parameters on the control group only (and thus impose the zero mean, unit variance on the control group latent distribution of θ_{id}), to allow for the fact that treatment status may alter the parameters of the model. As 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). 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.

4.2 Child Development

Child development is a multidimensional construct, as discussed, for instance, in Cunha et al. (2010), Attanasio, Meghir, and Nix (2017) and Attanasio et al. (2020b). 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 skills across different domains. The measures we used at endline were richer than those used at baseline, a choice driven by cost limitations and by the fact that the emphasis of the study is on estimating treatment effects on endline child development, with baseline measures primarily being useful to check for balance and to increase the precision of estimated effect sizes.

[17] We use Gauss-Hermite quadrature to approximate the integral over the unobserved latent factor.

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);[18] 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. ask parents 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.[19][20] For each assessment separately, we combine items using two-parameter IRT model described in Section 4.1.[21] Appendix Table C.3 presents the parameter estimates for these measurement models alongside estimated standard errors. 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.7), 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 from 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.[22] 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 used only 70 items, chosen through piloting. Inhibitory control, a dimension of executive functioning, was measured using the nonverbal

[18] We extended the ASQ for 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 the instrument had sufficient information over the entire support of baseline child development.

[19] For items belonging to the ASQ, we formally have three categories: "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).

[20] 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.

[21] The 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 and adding a dummy indicator for the assessment used.

[22] We 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

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 full 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 maternal 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, as in Bernal and Fernández (2013) and Andrew et al. (2018).[23]

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 skills are strongly correlated with age, baseline child development and household wealth 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 skills.

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 bootstrapped confidence intervals. Importantly, we notice that for both cognitive and socioemotional development, almost all of the items appear to be informative of the underlying factor. For cognitive development,

[23]This helps to ensure reliability and construct validity – the extent to which an instrument measures what it aims to measure – which can be challenging when translating and adapting across languages and cultures (see Peña, 2007).

for instance, all 152 of our estimated discrimination parameters (the β_{jd}'s) are positive and only 10 out of 152 have 90% 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 0.16 SD (0.14 SD) for cognitive development. The corresponding figures for socioemotional problems are 0.23 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 teachers' and teaching assistants' routines and the quality of their instruction. We first collected teachers' reported overtime hours measured as the number of hours they report working over and above their contracted hours on a typical week. We next move onto 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 with what frequency (in how many days) (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 "Caring Activities" which relate to basic care of children such as changing nappies, brushing teeth and washing hands, naps 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., 2020b; 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 "Caring 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 caring activities and adopt the graded-response specification described in equation (4.2). We repeat the identical procedure for the learning and development activities. Appendix Tables C.5 and C.6 show the full set of activities used and the estimated parameters in the measurement systems for teachers' learning and care activities respectively. Tables C.7 and C.8 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 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 across cognitive (Burchinal et al., 2000; Peisner-Feinberg et al., 2001) and social-emotional skills (Sylva et al., 2006). The ECERS-R was carried out by psychologists, who were trained for three weeks, and each observation lasted at least half of a school day. Due to logistical and budgetary constraints, we only conducted ECERS-R in 172 of the 847 classrooms in our sample.[24]

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).[25] [26] Appendix Table C.11 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

This section presents estimates of the impacts that HIM and HIM+FE had on child development. In Section 5.1, we present our estimates of the average impacts, as well as evidence on how the impacts differ by observed characteristics of the children and their families. In Section 5.2, we assess impacts on teacher behavior and directly observed teaching quality, which might be informative about the mechanisms at play. We then discuss these mechanisms further in Section 6, where we propose a conceptual framework through which to interpret our findings.

5.1 Effects on Child Development

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 development and socioemotional development into summary factors, as described in Section 4. Impacts on these factor scores have the advantage of

[24]The 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 it is too small to be analysed independently and cannot be linked to ECERS-R classroooms due to lack of common items.

[25]Each item is formed of around 10 sub-items grouped under the headings 'inadequate', 'minimal', 'good' and 'excellent' to which the observer must answer 'true' or 'false'. We followed the official administration procedure, which unfortunately turned out to be poorly suited to our context due to stopping rules which resulted in a high number of non-random missing values for items in the 'minimal', 'good' and 'excellent' categories. 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.

[26]To 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.

using information contained in each assessment more efficiently and providing impacts scaled by the variance of the underlying factor, uncontaminated by measurement error, in the control group.

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

5.1.1 Cognitive Skills

Columns 1–6 of Table 2 show estimates of the interventions' impacts 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	-0.007	1.914	-0.398	0.462	-1.238	0.300	0.635
	(0.410)	(2.836)	(1.275)	(0.763)	(1.387)	(0.372)	(2.606)
	p=0.987	*p=0.501*	*p=0.753*	*p=0.548*	*p=0.375*	*p=0.415*	*p=0.805*
	q=0.987	*q=0.944*	*q=0.979*	*q=0.930*	*q=0.920*	*q=0.935*	*q=0.972*
HIM+FE	0.910**	4.658*	2.426**	1.900***	0.743	0.352	-1.909
	(0.432)	(2.798)	(1.239)	(0.646)	(1.229)	(0.418)	(2.723)
	p=0.038	*p=0.085*	*p=0.049*	*p=0.004*	*p=0.544*	*p=0.394*	*p=0.494*
	q=0.188	*q=0.293*	*q=0.198*	*q=0.022*	*q=0.544*	*q=0.770*	*q=0.740*
Difference	0.917**	2.745	2.824**	1.438**	1.981*	0.053	-2.543
	(0.392)	(2.491)	(1.101)	(0.722)	(1.212)	(0.329)	(2.389)
	p=0.022	*p=0.294*	*p=0.012*	*p=0.040*	*p=0.097*	*p=0.858*	*p=0.288*
	q=0.105	*q=0.473*	*q=0.065*	*q=0.165*	*q=0.317*	*q=0.858*	*q=0.620*
N	1074	1074	1074	1075	1075	1075	1075
Control mean	486.307	464.309	460.128	49.841	33.541	7.142	58.300
Control SD	5.362	28.712	16.425	10.137	15.150	4.452	24.799

Note. Single-hypothesis two-sided p-values calculated using a block 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. (2016). 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 training program combined with the government improvement program (HIM+FE) did improve children's performance in several of the cognitive assessments. These estimated impacts are significantly different from zero at the 5% level for three assessments (fluid reasoning, expressive language and school readiness) and at the 10% level for the memory for words assessment. When examining the additional effect of FE over and above HIM ('Difference' row), we see statistically significant improvements across four measures. 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 0.17 of a standard deviation (SD) relative to the pure control, with a p-value of 0.010. The final row of Table 3 shows that the addition of the FE component resulted in a 0.15 SD improvement in child cognitive skills relative to HIM alone ($p = 0.009$).

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 or socioemotional 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.

5.1.2 Socioemotional Skills

The last column of Table 2 shows impacts on ASQ:SE (our measure of socioemotional problems - see Section 4.2 for more details) scored according to the publisher's guidelines, while 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. It should be noted that we may be under-powered to identify small impacts on this outcome - the larger standard errors in the socioemotional skills analysis (Table 3) indicate that these measures contain less information (see Section 4.2).

5.1.3 Robustness

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; it suggests that our results are not dependent on specific modelling choices we have made. In Appendix Table B.7, we show that our results are also not sensitive to the control variables we include. Even without any control variables, 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, 0.078 and 0.03. Controlling only for children's age leads of p-values of 0.035 and 0.015 respectively. Effect sizes and patterns of significance are left virtually unchanged with the addition of controls for gender, city and baseline child development (either as raw scores or as factor scores). Furthermore, Table B.8 shows that our findings are robust to constructing child development factor scores using a non-parametrically estimated distribution for the underlying latent factor, rather than imposing normality.

Table 3: Impacts on Cognitive and Socioemotional Factor Scores

	(1) Cognitive Development	(2) Socioemotional Problems
HIM	0.018	0.002
	(0.079)	(0.135)
	p=0.806	p=0.992
FE+HIM	0.168***	-0.126
	(0.066)	(0.143)
	p=0.010	p=0.389
Difference	0.150***	-0.129
	(0.058)	(0.143)
	p=0.009	p=0.381
N	1075	1056

*Note.*Single-hypothesis two-sided p-values calculated using a block 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. 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

Further, all our main conclusions hold when we also include younger children (below 48 months at endline) 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 find no evidence of heterogeneity by age within this extended sample (Appendix Table B.11).

5.1.4 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.[27] Estimates in column 1 of Table 4 show first that neither the wealthier nor the more disadvantaged children experienced improvements in cognitive

[27]This wealth index was constructed by summarizing information about whether the household owned at least one of 18 different assets (including, for example, a car, a TV or a washing machine) through factor analysis.

development as the result of the HIM program. In contrast, the HIM+FE program had a relatively large impact of 0.31 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).

Table 4: Heterogeneity by Wealth and Baseline Child Development

	Panel A: Cognitive		***Panel B: Socio-Emotional***	
	(1)	(2)	(3)	(4)
HIM	0.079	0.103	0.095	0.155
	(0.106)	(0.122)	(0.180)	(0.188)
	p=0.459	*p=0.409*	*p=0.581*	*p=0.409*
HIM+FE	0.307***	0.268***	-0.050	-0.097
	(0.089)	(0.090)	(0.169)	(0.188)
	p=0.000	*p=0.004*	*p=0.754*	*p=0.596*
HIM x Wealthier	-0.118		-0.186	
	(0.121)		(0.187)	
	p=0.332		*p=0.337*	
HIM+FE x Wealthier	-0.266**		-0.162	
	(0.121)		(0.209)	
	p=0.029		*p=0.423*	
HIM x Higher BL dev		-0.175		-0.284
		(0.136)		(0.228)
		p=0.207		*p=0.196*
HIM+FE x Higher BL dev		-0.202		-0.059
		(0.131)		(0.244)
		p=0.130		*p=0.788*
Difference	0.227***	0.164*	-0.145	-0.252
	(0.076)	(0.090)	(0.169)	(0.188)
	p=0.001	p=0.071	p=0.380	p=0.188
Difference x Wealthier	*-0.148*		*0.024*	
	(0.097)		*(0.209)*	
	p=0.123		*p=0.915*	
Difference x Higher BL dev		-0.026		0.225
		(0.119)		(0.244)
		p=0.815		*p=0.341*
N	1075	1075	1056	1056

Note. 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, 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 5.1.5.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.1.5 Heterogeneity by Baseline Development

The evidence on heterogeneity by baseline child development is less strong. 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".[28] We find that the point estimate of the impact of HIM+FE on children with lower-than-median baseline development, at 0.27 SD (reported in column 2 of Table 4), is higher than the average impact; however, the impact on children with a higher level of development is not statistically different. The impacts on socioemotional development are not different from zero for either group (column 4).

5.2 Mechanisms

We now turn to exploring the impacts of the two interventions on a number of proxies for potential mechanisms. We look at the impacts of the interventions on reported weekly hours of teacher overtime as a proxy for impacts on total time spent on their job. We then look at impacts on more detailed measures of teacher behavior that capture how teachers allocate their time in the classroom. We use data from the Teacher Survey of Early Education Quality (TSEEQ) to construct a "Learning Activities" factor from data on frequency with which teachers reported conducting various learning activities in the classroom, as well as a "Caring Activities" factor using data on frequency with which they reported conducting personal care activities (see Section 4 for details). To analyze how the addition of FE changed TAs' routines we construct the same measures using TAs' responses to these same questions. Finally, we use the ECERS-R data to construct a measure of quality of teaching within the classroom, as observed by trained psychologists.

First, however, we ask how all of these measures of classroom environment are correlated with children's development. It is important to note that we are simply examining correlations here, we are not identifying the causal effect of these inputs on child development. In Table 5, we report the results of a regression of the child cognitive development factor (used in our main impact analysis in Table 4) on indicators of how much overtime teachers reported, on the factors capturing caring and learning activities, and on the quality of classroom teaching directly observed during the ECERS-R assessment. All these indicators are 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. While columns 1 and 3 show a positive and significant association between child cognitive development and teacher-reported learning and development activities in the classroom and overtime, there is 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

[28] Specifically, we age-standardize all sub-scales of the ASQ-3 from baseline as well as the MacArthur-Bates CDI 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 5: Correlations between Child Cognitive Development and Teacher-Reported Activities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Learning Activities	0.073**				0.090*	0.079	0.057	0.06
	(0.036)				(0.052)	(0.049)	(0.051)	(0.093)
	p=0.045				*p=0.086*	*p=0.105*	*p=0.270*	*p=0.519*
Caring Activities		0.009			-0.025	-0.024	-0.012	-0.033
		(0.039)			(0.049)	(0.048)	(0.050)	(0.094)
		p=0.827			*p=0.618*	*p=0.617*	*p=0.814*	*p=0.729*
Overtime			0.049*			0.042*	0.025	0.063
			(0.026)			(0.026)	(0.030)	(0.076)
			p=0.066			*p=0.100*	*p=0.404*	*p=0.415*
Observed Teaching Quality (ECERS-R)				0.215*			0.198*	0.329*
				(0.123)			(0.114)	(0.186)
				p=0.085			*p=0.085*	*p=0.087*
N	1075	1075	1075	727	1075	1075	727	249

Note. Single-hypothesis 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

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 much less precise. This pattern of a significant association between child development and learning activities / good teaching processes but not care activities is consistent with the message of the FE program, which emphasized the importance of teachers focusing on learning rather than care activities.

Having established that they are correlated child development, next we consider the impact of the two programs on teacher and TA activities, as well as quality of classroom processes. We find substantial differences in the ways that the two programs affected these (Table 6). 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 0.35 SD reduction in frequency of learning activities and a 0.85 SD reduction in the frequency of personal care activities. Overall, teacher-reported overtime also fell by nearly half an hour per week (relative to 1.2 hours among teachers in the control group).

Further, we find that combining the introduction of TAs with the FE teacher training program had a different impact on teachers' routines than introduction of TAs without the training. Table 6 shows that while we still observe a reduction in personal care activities relative to the control group of roughly the same size as in the HIM arm, there is no discernible effect of the program on learning activities or overtime. If we compare HIM+FE with HIM, we see that the addition of FE offset the reduction in the time teachers dedicate to learning activities as well as the reduction in overtime induced by HIM alone.[29] It did not,

[29]Our results suggest that FE increased the amount of effort that teachers in a given HI allocated to classroom teaching and to learning activities in particular. It is most natural to think that FE increased the effort of the teachers who were already employed by the HI. It might be the case that FE could 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

Table 6: Impacts on Teachers' and TAs' Behavior

	Teachers' time	***Teachers' activities***		***TA's activities***		***Observed Teaching***
	Overtime	Learning	Care	Learning	Care	***Quality***
	(1)	(2)	(3)	(4)	(5)	(6)
HIM only	-0.390*	-0.350**	-0.845**			-0.030
	(0.212)	(0.151)	(0.325)			(0.106)
	p=0.062	*p=0.025*	*p=0.017*			*p=0.727*
FE+HIM	0.029	-0.054	-0.721*			0.196**
	(0.280)	(0.152)	(0.363)			(0.094)
	p=0.940	*p=0.702*	*p=0.062*			*p=0.042*
Difference	0.418	0.296**	0.124	0.235	0.017	0.226**
	(0.280)	(0.143)	(0.363)	(0.209)	(1.879)	(0.104)
	p=0.128	*p=0.037*	*p=0.733*	*p=0.268*	*p=0.942*	*p=0.040*
N	841	841	841	254	254	172

Note. Single-hypothesis two-sided p-values calculated using a block bootstrap, resampling triplets with replacement (1,000 iterations). 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 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

however, alter the reduction in the time teachers dedicate to care activities that HIM brought about, nor did it have an impact on TA time-allocation (Columns 4 and 5).

Finally, in Column 6, we show the impacts on the ECERS-R measure - psychologists' assessment of overall quality of teaching in the classroom. Teachers were always present in the classroom during the assessment, therefore this is a proxy of the quality of teachers' teaching rather than the quantity. We only have these measures for 172 out of the 841 classrooms in the sample (see Section 4.3 for details). We see no evidence that HIM affected the quality of the learning environment, a finding that is consistent with null effects on child development. Even for this small sample, however, we do see evidence that the addition of FE was effective at improving the quality of teaching processes. Compared to the pure control, we estimate that the HIM+FE program improved the quality of directly-observed teaching by 0.20 SD ($p = 0.042$), with the difference between the HIM and HIM+FE arms being similar in magnitude and statistical significance. This evidence is consistent with the fact that this measure is intended to capture quality of pedagogy and teacher–child interaction, which were the areas targeted by the FE program.

Taken together, these results suggest that teachers' behavioral reactions are key 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 their importance 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 adapt their own activities in the classroom in a way that was conducive to improvements in children's development.

were employed in the center at baseline (see Table B.12).

In the next section, we propose a model that could explain the differential impact of the two interventions on child development through mechanisms that are consistent with this narrative and sequence of impacts of the two programs.

6 Interpreting the Results: A Conceptual Framework

In this section we develop a model which is useful for considering how providing preschool teachers with help in the classroom and with information might affect teachers' behavior and child development. 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 teachers may misperceive the process of child development in order to consider the effects of combining the additional resources that teachers receive with teacher training. We model three types of misperceptions that we think may have been corrected by the FE curriculum (see Section 2). These include misperceptions 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.

The model we propose 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 within which to interpret our results rather than unambiguous predictions about the impacts of these different programs.

6.1 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:

$$1 = K + N = K + L_t + C_t = K + \tau_t N + (1 - \tau_t)N \tag{6.1}$$

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:

$$H = z f(L_t, C_t, A) \tag{6.2}$$

where z denotes total factor productivity. 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.[30] 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 depend positively on the amount of time they do *not* spend in the classroom and thus can spend on leisure and/or administration, K.

$$u(H, K) = u(H, 1 - L_t - C_t) \tag{6.3}$$

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 $u(.,.)$ 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:

$$\max_{L_t, C_t} u(zf(L_t, C_t, A), 1 - L_t - C_t) \tag{6.4}$$

subject to the production function in equation (6.2). This gives two first-order conditions:

$$0 = zu_H f_1 - u_K$$

$$0 = zu_H f_2 - u_K$$

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

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

6.2 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

[30] It 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).

activities. Differentiating the first-order conditions for L_t and C_t with respect to A and rearranging, we get:

$$\frac{dL_t^*}{dA} = \frac{f_{12} - f_{22}}{X}\left[\underbrace{zu_{HH}\frac{u_K}{zu_H}f_3}_{(1)} + \underbrace{zu_H\left(f_{13} + f_{12}\frac{f_{23} - f_{13}}{f_{12} - f_{22}}\right)}_{(2)} + \underbrace{\frac{f_{13} - f_{23}}{f_{12} - f_{22}}\left(-zu_{HH}\left(\frac{u_K}{zu_H}\right)^2 - u_{KK}\right)}_{(3)}\right] \tag{6.5}$$

$$\frac{dC_t^*}{dA} = \frac{f_{12} - f_{11}}{X}\left[\underbrace{zu_{HH}\frac{u_K}{zu_H}f_3}_{(1)} + \underbrace{zu_H\left(f_{23} + f_{12}\frac{f_{13} - f_{23}}{f_{12} - f_{11}}\right)}_{(2)} + \underbrace{\frac{f_{23} - f_{13}}{f_{12} - f_{11}}\left(-zu_{HH}\left(\frac{u_K}{zu_H}\right)^2 - u_{KK}\right)}_{(3)}\right] \tag{6.6}$$

where $X > 0$ (see Appendix D.1 for expression). We consider first the overall effect of an increase in TA time on the amount of time teachers spend on learning activities. Equation (6.5) shows this overall effect is comprised of three effects, numbered (1) to (3):

1. A **resource, or income, 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.[31]

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 (6.6). The total amount of time teachers spend on classroom activities will only depend on the income and complementarity effects:

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

[31] We 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.

6.3 Impacts of Teacher Training

The second intervention, HIM+FE, combined the introduction of TAs with a training program for existing teachers. In our framework, training could have several impacts. It could improve the total factor productivity with which inputs are translated into child development. This would correspond to an increase in the z term in equation (6.2) above. However, as we discuss below, it is unclear how such an increase alone could produce the patterns of changes in teachers' time-use that we observed in response to the introduction of FE to the HIM program (Table 6).

A different possibility is that, before the training, teachers might have had misperceptions over one or more aspects of the process of child development, including the relative importance of different types of activities in the classroom for child development and the substitutability between TA inputs and their own. For example, if teachers underestimate the productivity of doing learning activities with their class, they may dedicate less of their time to these activities than they would under full information. The training could improve teachers' understanding of child development.

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 (a) training improves true or perceived total factor productivity of activities performed by teachers and TAs; (b) training corrects teachers' misperceptions about the relative importance of learning and care activities for child development; and (c) training corrects teachers' misperceptions about the degree of substitutability between them and TAs in the performance of different types of activities in the classroom.

6.3.1 Introducing Distorted Perceptions

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

$$A = L_a + C_a = \tau_a A + (1 - \tau_a)A \tag{6.7}$$

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.

We work with a specific production function for child development which is a special case of that considered above. In particular, we assume 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. We allow for the possibility that teachers

have misperceptions about the process of child development but we do not allow these to be completely arbitrary. In particular, following a similar approach to that used by Caucutt et al. (2017) in a different context, we allow for misperceptions regarding the level of total productivity, the weight given to learning activities and the relative productivity of teachers and TAs. In what follows, symbols with a $\tilde{}$ denote *perceived* parameters, corresponding to the *true* parameters which do not have a $\tilde{}$. We assume that the perceived process of child development, on the basis of which teachers decide how to allocate their time, has the same form as the true process, but with possibly biased parameters.

$$H = \tilde{z}\left(\tilde{w}^{1-\rho}\tilde{L}^{\rho} + (1-\tilde{w})^{1-\rho}\tilde{C}^{\rho}\right)^{\frac{1}{\rho}} \tag{6.8}$$

where $0 < \tilde{z} \leq z$ is perceived total factor productivity and $0 < \tilde{w} \leq w$ represent the teachers' perceptions of the true parameters w, which determines 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:

$$\begin{aligned} \tilde{L} = &\quad \left(\tilde{\theta}_l L_t^{\lambda} + (1-\tilde{\theta}_l)L_a^{\lambda}\right)^{\frac{1}{\lambda}} = \left(\tilde{\theta}_l \tau_t^{\lambda} N^{\lambda} + (1-\tilde{\theta}_l)\tau_a^{\lambda} A^{\lambda}\right)^{\frac{1}{\lambda}} \qquad \lambda \in (0,1] \\ \tilde{C} = &\quad \left(\tilde{\theta}_c C_t^{\lambda} + (1-\tilde{\theta}_c)C_a^{\lambda}\right)^{\frac{1}{\lambda}} = \left(\tilde{\theta}_c (1-\tau_t)^{\lambda} N^{\lambda} + (1-\tilde{\theta}_c)(1-\tau_a)^{\lambda} A^{\lambda}\right)^{\frac{1}{\lambda}} \end{aligned} \tag{6.9}$$

where $\tilde{\theta}_l$ and $\tilde{\theta}_c$ represent teachers' perceptions about θ_l and θ_c – true relative efficiency of teachers and TAs at performing learning and care activities, respectively. We note that the parameters of the aggregator functions in equation (6.9) 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.[32]

The specification in equations (6.8) and (6.9) thus allows for the possibility that teachers misperceive total factor productivity (z), the importance of learning activities (w) and the relative productivity of TAs (θ_l and θ_c).[33]

6.3.2 Teachers' Decision Problem under Distorted Perceptions

In this specification of the model, teachers choose three variables, L_t, C_t and L_a,[34] 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

[32] The 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.

[33] As noted above, the specification in equations (6.8) and (6.9) can be considered a special case of equation (6.2), although this last equation does not consider the allocation of TA time to specific activities. One could obtain a specification like equation (6.2) by assuming that TA time allocation to learning activities, τ_a, is chosen optimally, given A and the parameters of the production function.

[34] As A is given exogenously, a choice of L_a determines C_a.

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:

$$\tau_t: \quad 0 = \frac{\tilde{z}}{\rho} h^{\frac{1}{\rho}-1} \left(\tilde{w}^{1-\rho} \frac{\partial \bar{L}^\rho}{\partial \tau_t} + (1-\tilde{w})^{1-\rho} \frac{\partial \bar{C}^\rho}{\partial \tau_t} \right) \tag{6.10}$$

$$\tau_a: \quad 0 = \frac{\tilde{z}}{\rho} h^{\frac{1}{\rho}-1} \left(\tilde{w}^{1-\rho} \frac{\partial \bar{L}^\rho}{\partial \tau_a} + (1-\tilde{w})^{1-\rho} \frac{\partial \bar{C}^\rho}{\partial \tau_a} \right) \tag{6.11}$$

where $h = \left(\tilde{w}^{1-\rho} \tilde{L}^\rho + (1-\tilde{w})^{1-\rho} \tilde{C}^\rho \right)$. This last term, as well as $\tilde{z}$, cancels out from both first-order conditions. Substituting in equation (6.10) the expressions for $\partial \bar{L}/\partial \tau_t$ and $\partial \bar{C}/\partial \tau_t$, it is straightforward to obtain:

$$\left(\frac{\tilde{w}}{1-\tilde{w}} \right)^{1-\rho} \frac{\tilde{\theta}_l}{\tilde{\theta}_c} = \left(\frac{h_c}{h_l} \right)^{\frac{\rho}{\lambda}-1} \left(\frac{\tau_t}{1-\tau_t} \right)^{1-\lambda} \tag{6.12}$$

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 (6.11), we obtain:

$$\left(\frac{\tilde{w}}{1-\tilde{w}} \right)^{1-\rho} \frac{1-\tilde{\theta}_l}{1-\tilde{\theta}_c} = \left(\frac{h_c}{h_l} \right)^{\frac{\rho}{\lambda}-1} \left(\frac{\tau_a}{1-\tau_a} \right)^{1-\lambda} \tag{6.13}$$

From these expressions, it is straightforward to see that when there are no TAs, we get:

$$\frac{\tilde{w}}{1-\tilde{w}} = \frac{\tau_t}{1-\tau_t} \tag{6.14}$$

Taking the ratio of equations (6.13) and (6.12), we obtain:

$$\frac{\tilde{\theta}_l}{1-\tilde{\theta}_l} \frac{1-\tilde{\theta}_c}{\tilde{\theta}_c} = \left(\frac{\tau_t}{\tau_a} \right)^{1-\lambda} \left(\frac{1-\tau_a}{1-\tau_t} \right)^{1-\lambda} \tag{6.15}$$

6.3.3 Implications for HIM+FE Program Impacts

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 correcting $\tilde{z}$, $\tilde{\theta}_l$, $\tilde{\theta}_c$ and $\tilde{w}$ on teacher and TA behavior through our model.

1. **Correcting $\tilde{z}$:** As noted above, z cancels out from both first-order conditions in equations (6.10) and (6.11). 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. **Correcting $\tilde{w}$:** 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, an increase in $\tilde{L}^*$, which is determined both by τ_t and τ_a. However, as we have seen in equation (6.15), the ratio of τ_t and τ_a does not depend on $\tilde{w}$ and, therefore, will stay constant. It follows that *an exogenous increase in the perceived importance and productivity of learning activities, relative to care activities, will result in both teachers and TAs spending a greater proportion of their time on learning activities.* In other words, both τ_t and τ_a will increase as a consequence of an increase in $\tilde{w}$ (See Appendix D.2.2 for details). 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.

3. **Correcting $\tilde{\theta}_l$ and/or $\tilde{\theta}_c$:** From equation (6.15), 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.

6.4 What We Learn from the Model

Our empirical results suggest that the introduction of TAs through HIM had no impact on child development and results in a considerable reduction in the amount of time that teachers spend on learning and caring activities in the classroom (and a reduction in overtime). This suggests that the introduction of TAs led to a negative resource effect and/or a negative substitutability effect on teachers' input in the classroom, as set out in Section 6.2.[35]

The addition of teacher training in the form of FE generated a substantial positive impact on child development alongside a significant change in teacher time-allocation relative to HIM alone. This is despite the fact that teachers in HIM and HIM+FE preschools had access to the exact same resources, including TAs. As shown in Section 6.3, an increase in total factor productivity or perceived total factor productivity alone could not generate these changes. However, *on their own*, changes in teachers' beliefs about the relative importance of learning and care activities for child development or about how productive they are relative to TAs in performing learning and care activities also could not generate the observed changes. This is because, while both are consistent with an increase in time that teachers spend on learning activities that we see, neither is consistent with no change in TA time-use. However, if both are at play then the lack of effects on TA time could be explained by the effects of the two changes canceling each other out.[36] Therefore, a

[35] The results reported in equation (6.5) and (6.6) also hold in the version of the model we use in Section 6.3.

[36] An increase in $\tilde{\theta}_l$ relative to $\tilde{\theta}_c$ would also imply that teachers would spend a smaller share of time on care activities than

plausible narrative is that the FE teacher training program increased both the perception of the importance of learning activities and the perception of teachers' comparative advantage in learning activities, as well as, possibly, total factor productivity.

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 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 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 31% 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, at 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. First we show that provision of TAs resulted in teachers reducing their time at work, including on learning activities that are highly correlated with child development. However, the addition of the teacher training program induced teachers to increase time spent at work, including on learning activities. We then consider these and our main impact results through the lens of a theoretical model. This allows us to show that the zero impact of hiring TAs can be generated by the interaction of three effects: a resource effect, a substitutability/complementarity effect, and a comparative advantage effect. Furthermore, the teacher time-use response is consistent with the teacher training program successfully correcting teachers' misperceptions about the process of child development.

TAs which, since we see an increase in teacher time on learning activities and no change in time spent on care activities, would be achieved through a decrease in TA time on learning activities and an increase in TA time on care activities. In contrast, an increase in $\tilde{w}$ would imply an increase in time TAs spend on learning activities. Since TA time is fixed, achieving this increase would necessitate a reduction in TA time on care activities. Thus the combined effect of increases in $\tilde{\theta}_l$ and $\tilde{w}$ works in the same direction for teachers but in opposite directions for TAs and is hence consistent with our pattern of results: an increase in the amount of time teachers spend on learning activities and no change in TA time-allocation.

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, viewed through the lens of our model, 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, but happened in the HIM+FE program. 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." .

Agostinelli, Francesco and Matthew Wiswall. 2016. "Estimating the Technology of Children's Skill Formation." *NBER Working Paper* (July).

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, Teodora Boneva, and Christopher Rauh. 2020a. "Parental Beliefs about Returns to Different Types of Investments in School Children." *Journal of Human Resources* :0719–10299R1.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020b. "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia." *American Economic Review* 110 (1):48–85.

Attanasio, Orazio, Flávio Cunha, and Pamela Jervis. 2019. "Subjective Parental Beliefs. Their Measurement and Role." *NBER Working Paper* :No. 26516.

Attanasio, Orazio, Costas Meghir, and Emily Nix. 2017. "Human Capital Development and Parental Investment in India."

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* .

Barrera-Osorio, Felipe, Kathryn Gonzalez, Francisco Lagos, and David J. Deming. 2020. "Providing performance information in education: An experimental evaluation in Colombia." *Journal of Public Economics* 186:104185.

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).

Bergman, Peter. 2021. "Parent-child information frictions and human capital investment: Evidence from a field experiment." *Journal of Political Economy* 129 (1):286–322.

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.

Boneva, Teodora and Christopher Rauh. 2018. "Parental Beliefs about Returns to Educational Investments—The Later the Better?" *Journal of the European Economic Association* 16 (6):1669–1711.

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." 25 (2).

Burchinal, Margaret R, Joanne E Roberts, Rhodus Riggins Jr, Susan A Zeisel, Eloise Neebe, and Donna Bryant. 2000. "Relating quality of center-based child care to early cognitive and language development longitudinally." *Child development* 71 (2):339–357.

Carneiro, Pedro Manuel, Emanuela Galasso, Italo Lopez Garcia, Paula Bedregal, and Miguel Cordero. 2021. "Parental Beliefs, Investments, and Child Development: Evidence from a Large-Scale Experiment." *SSRN Electronic Journal* (February).

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, Flávio, Irma Elo, and Jennifer Culhane. 2013. "Eliciting Maternal Expectations about the Technology of Cognitive Skill Formation." 4 (1):1–23.

———. 2020. "Maternal subjective expectations about the technology of skill formation predict investments in children one year later." *Journal of Econometrics* .

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3):883–931.

Danzer, Virginia A, Mary Frances Gerber, Theresa M Lyons, and Judith K Voress. 1991. *Daberon 2: Screening for School Readiness*. 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.

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.

Dizon-Ross, Rebecca. 2019. "Parents' beliefs about their children's academic ability: Implications for educational investments." *American Economic Review* 109 (8):2728–2765.

Duflo, Annie, Jessica Kiessel, and Adrienne M Lucas. 2020. "External Validity: Four Models of Improving Student Achievement." .

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." Tech. rep., National Bureau of Economic Research, Cambridge, MA.

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

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."

Havnes, Tarjei and Magne Mogstad. 2015. "Is universal child care leveling the playing field?" *Journal of Public Economics* 127:100–114.

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 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.

Kinsler, Josh and Ronni Pavan. 2021. "Local distortions in parental beliefs over child skill." *Journal of Political Economy* 129 (1):81–100.

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, Azeem Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." .

Neuman, Michelle J and Lynette Okeng'o. 2019. "Early childhood policies in low-and middle-income countries."

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.

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

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." Tech. rep.

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 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 pedagogical training program, which we argue was the key component of FE in terms of generating child development impacts. FE provided us with the total cost of this component (COP 419546284, or USD 233081 at February 2013 exchange rate of 1800 COP/USD). With this budget, FE provided completed training for 99 teachers from the 40 HIs in the HIM+FE treatment arm.[37] This represented 32% of teachers who worked in those 40 HIs. We thus estimate that the initial one-off cost to roll out the pedagogical training program to new HIs, at the same intensity as achieved the impacts we see in this study (i.e. training 32% of teachers), would be USD 5827 per HI or USD 35 per child attending an HI.

However, it is unreasonable to assume that the same intensity of training program would be required year after year for FE to sustain its impacts on successive cohorts. Rather, we calculate the costs of maintaining a ratio of training 32% of staff, which implies providing training for 32% of new staff, and the costs of providing a yearly refresher training to all teachers who have already been trained which we assume would cost 25% of the costs of the full training. Given these assumptions, we estimate that the ongoing cost per center of maintaining the results of the pedagogical training program would be USD 2206 per year and the cost per child would be USD 13 per year. All data, assumptions and formulae used in these calculations are shown in Table 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 they teach, and the cost per child figure is based on the total number of children in the center. Our study children were between 1 and 3 at baseline and between 3 and 5 at endline. Given only 2.7% of teachers report that they primarily teach children younger than one year, we consider the training of all teachers relevant for generating the treatment effect. Moreover, we note that teachers' propensity to complete the training appears independent of the age of the children they teach. Therefore, we include all teachers and children of all ages in the costing.

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

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

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

B Additional Tables

Table B.1: Attrition by Treatment Status

	Extended Sample	Analysis Sample
	All kids	Kids 48 months +
HIM vs. control	0.008	-0.014
	(0.020)	(0.017)
	p=0.695	*p=0.422*
FE+HIM vs. control	0.015	-0.013
	(0.019)	(0.018)
	p=0.412	*p=0.468*
Difference	0.008	0.001
	(0.017)	(0.017)
	p=0.658	*p=0.420*
N	1987	1149

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Standard errors (clustered at the triplet level) in parentheses for 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.

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***	0.992***	0.927***	0.688***	1.137***	0.542***	0.439***
	(0.049)	(0.060)	(0.135)	(0.098)	(0.124)	(0.070)	(0.035)
	p<0.001	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*
FE+HIM	0.871***	0.983***	0.756***	0.627***	1.014***	0.506***	0.410***
	(0.045)	(0.050)	(0.151)	(0.089)	(0.129)	(0.081)	(0.022)
	p<0.001	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*	*p<0.001*
Difference	0.008	-0.010	-0.171	-0.060	-0.123	-0.036	-0.029
	(0.058)	(0.070)	(0.135)	(0.076)	(0.135)	(0.065)	(0.038)
	p=0.891	*p=0.890*	*p=0.205*	*p=0.430*	*p=0.364*	*p=0.586*	*p=0.448*
N	120	120	120	120	120	120	120
Control mean	0.073	0.084	0.552	0.291	0.319	0.11	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. Single-hypothesis two-sided *p*-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations). 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	Wealth index	Mother's education	N
Fluid Reasoning	0.164***	0.182***	0.226***	0.074**	0.073**	0.101***	0.253***	0.149***	0.192***	1,074
Memory for Words	0.193***	0.129***	0.198***	0.091***	0.114***	0.088***	0.218***	0.074**	0.096***	1,074
Expressive Language	0.163***	0.187***	0.214***	0.081***	0.059*	0.081***	0.240***	0.224***	0.273***	1,074
School Readiness	0.283***	0.211***	0.290***	0.092***	0.122***	0.159***	0.298***	0.166***	0.210***	1,075
Receptive Language	0.221***	0.202***	0.217***	0.088***	0.073**	0.117***	0.250***	0.200***	0.248***	1,075
Inhibitory Control	0.192***	0.090***	0.165***	0.022	0.061**	0.086***	0.156***	0.101***	0.120***	1,075
Socio-Emotional Problems	-0.035	-0.132***	-0.133***	-0.036	-0.062**	-0.149***	-0.147***	-0.150***	-0.140***	1,075

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	0.270***	0.317***	0.289***	0.477***	0.334***	1	
Socio-Emotional Problems	-0.156***	-0.0881***	-0.216***	-0.221***	-0.225***	-0.147***	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	0.018	0.002
	(0.079)	(0.135)
	[p=0.806]	[p=0.992]
HIM+FE	0.168***	-0.126
	(0.066)	(0.143)
	[p=0.010]	[p=0.389]
BL MacArthur Bates (older kids)	0.183***	-0.102**
	(0.038)	(0.053)
	[p=0.000]	[p=0.049]
BL MacArthur Bates (younger kids)	0.133*	-0.097
	(0.075)	(0.094)
	[p=0.062]	[p=0.302]
=1 if older	-0.016	0.037
	(0.131)	(0.175)
	[p=0.904]	[p=0.838]
BL ASQ Communication	0.248***	-0.161**
	(0.048)	(0.059)
	[p=0.000]	[p=0.012]
BL ASQ Gross Motor	-0.006	0.024
	(0.049)	(0.051)
	[p=0.917]	[p=0.629]
BL ASQ Fine Motor	-0.034	0.012
	(0.033)	(0.053)
	[p=0.321]	[p=0.828]
BL ASQ Problem Solving	0.179***	-0.037
	(0.049)	(0.074)
	[p=0.000]	[p=0.611]
BL ASQ Socio-Individual	-0.014	-0.136*
	(0.055)	(0.072)
	[p=0.800]	[p=0.055]
BL ASQ Socioemotional	0.031	0.228***
	(0.042)	(0.062)
	[p=0.440]	[p=0.000]
Male	-0.005	0.075
	(0.053)	(0.070)
	[p=0.915]	[p=0.273]
Age in Months	-0.833	0.933
	(0.669)	(1.013)
	[p=0.140]	[p=0.237]
Age in Months squared	0.009	-0.009
	(0.006)	(0.010)
	[p=0.118]	[p=0.228]
N	1075	1056
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. Single-hypothesis two-sided p-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations) and re-estimating the measurement system on each bootstrap. Standard errors (bootstrapped) in parentheses.

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

Table B.7: Sensitivity of Estimates to Different Control Variables

	Panel A: Cognitive						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIM	0.007	0.019	0.029	0.023	0.015	0.018	-0.017
	(0.093)	(0.089)	(0.088)	(0.085)	(0.087)	(0.079)	(0.081)
	p=0.949	*p=0.830*	*p=0.749*	*p=0.804*	*p=0.863*	*p=0.806*	*p=0.805*
FE+HIM	0.137*	0.159**	0.166**	0.155**	0.161**	0.168***	0.148**
	(0.079)	(0.078)	(0.077)	(0.073)	(0.068)	(0.066)	(0.065)
	p=0.078	*p=0.035*	*p=0.026*	*p=0.026*	*p=0.015*	*p=0.010*	*p=0.017*
Difference	0.130**	0.139**	0.136**	0.132***	0.146**	0.150***	0.166***
	(0.060)	(0.057)	(0.056)	(0.050)	(0.067)	(0.058)	(0.060)
	p=0.030	*p=0.015*	*p=0.015*	*p=0.009*	*p=0.026*	*p=0.009*	*p=0.005*
N	1075	1075	1075	1075	1075	1075	1075
Controls							
Age		X	X	X		X	X
Gender			X	X		X	X
City				X		X	X
Baseline Child Development (factor scores)					X	X	
Baseline Child Development (raw scores)							X

	Panel B: Socio-Emotional						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
HIM	-0.050	-0.058	-0.074	-0.046	-0.020	0.002	0.046
	(0.125)	(0.124)	(0.124)	(0.128)	(0.133)	(0.135)	(0.128)
	p=0.686	*p=0.649*	*p=0.548*	*p=0.721*	*p=0.887*	*p=0.992*	*p=0.726*
FE+HIM	-0.146	-0.151	-0.163	-0.124	-0.161	-0.126	-0.094
	(0.138)	(0.137)	(0.139)	(0.139)	(0.141)	(0.143)	(0.138)
	p=0.298	*p=0.267*	*p=0.232*	*p=0.388*	*p=0.271*	*p=0.389*	*p=0.496*
Difference	-0.096	-0.094	-0.088	-0.078	-0.141	-0.129	-0.140
	(0.138)	(0.137)	(0.139)	(0.139)	(0.141)	(0.143)	(0.138)
	p=0.498	*p=0.498*	*p=0.529*	*p=0.575*	*p=0.336*	*p=0.381*	*p=0.331*
N	1056	1056	1056	1056	1056	1056	1056
Controls							
Age		X	X	X		X	X
Gender			X	X		X	X
City				X		X	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. Single-hypothesis two-sided *p*-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations) and re-estimating the measurement system on each bootstrap. Standard errors (bootstrapped) in parentheses.
* p < 0.1, ** p < 0.05, *** p < 0.01

Table B.8: Sensitivity of Estimates to Relaxing Normality Assumption on Underlying Factors

	(1) Cognitive development	(2) Socioemotional problems
HIM	0.04	0.036
	(0.087)	(0.154)
	p=0.650	*p=0.818*
FE+HIM	0.200**	-0.167
	(0.082)	(0.187)
	p=0.019	*p=0.377*
Difference	0.160**	-0.202
	(0.071)	(0.151)
	p=0.030	*p=0.187*
N	1075	1063

Note. Table shows impacts on cognitive and socioemotional factors estimated without imposing normality on the underlying distribution, but instead using an empirical histogram described by Bock and Aitkin (1981). Due to computational intensity, we do not bootstrap these alternative measurement systems. Instead standard errors and p-values (clustered at the triplet level) here are estimated analytically. 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.9: Impacts on Externally-Standardized Scores in the Extended Sample

	Fluid Reasoning	Expressive Language	School Readiness
HIM only	0.218	-0.649	0.226
	(0.404)	(1.113)	(0.689)
	p=0.599	*p=0.547*	*p=0.723*
HIM+FE	0.964**	1.978*	1.210**
	(0.410)	(1.043)	(0.618)
	p=0.017	*p=0.062*	*p=0.048*
Difference	0.746**	2.627***	0.984
	(0.315)	(0.817)	(0.629)
	p=0.016	*p=0.001*	*p=0.118*
N	1837	1833	1835
Control mean	484.781	457.002	45.949
Control SD	6.516	16.887	11.435

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. Single-hypothesis two-sided p-values calculated using block bootstraps, 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. 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	0.018	0.013	-0.005	0.002	0.056	0.054
	(0.079)	(0.066)	(0.046)	(0.135)	(0.113)	(0.070)
	p=0.806	p=0.829	p=0.900	p=0.992	p=0.624	p=0.425
FE+HIM	0.168***	0.125**	-0.042	-0.126	-0.094	0.033
	(0.066)	(0.059)	(0.044)	(0.143)	(0.131)	(0.082)
	p=0.010	p=0.030	p=0.310	p=0.389	p=0.467	p=0.699
Difference	0.150***	0.112**	-0.037	-0.129	-0.149	-0.021
	(0.058)	(0.059)	(0.037)	(0.143)	(0.131)	(0.076)
	p=0.009	p=0.049	p=0.306	p=0.381	p=0.252	p=0.776
N	1075	1839		1056	1815	

Note. Table shows impacts in the main analysis sample (also shown in Table 3) and the extended sample 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. Single-hypothesis two-sided p-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations) and re-estimating the measurement system on each bootstrap. 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	0.013	0.13
	(0.112)	(0.140)
	p=0.907	p=0.357
(b) HIM x Middle	0.05	-0.081
	(0.088)	(0.128)
	p=0.574	p=0.529
(c) HIM x Oldest	-0.019	0.023
	(0.076)	(0.154)
	p=0.808	p=0.880
(d) HIM+FE x Youngest	0.107	0.013
	(0.106)	(0.139)
	p=0.318	p=0.926
(e) HIM+FE x Middle	0.142	-0.037
	(0.093)	(0.157)
	p=0.136	p=0.812
(f) HIM+FE x Oldest	0.133*	-0.216
	(0.072)	(0.184)
	p=0.073	p=0.247
N	1839	1800
p-value for testing		
(a)=(b)=(c)	0.737	0.388
(d)=(e)=(f)	0.963	0.424

Note. Table shows estimates of heterogeneous treatment effects by age in the extended sample 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. Standard errors in parentheses and p-values allow for clustering at the triplet level. 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

	Teachers' time Overtime	***Teachers' activities*** Learning	Care
	(1)	(2)	(3)
HIM only	-0.504*	-0.344**	-0.879**
	(0.271)	(0.140)	(0.340)
	p=0.062	*p=0.014*	*p=0.021*
FE+HIM	0.174	0.020	-0.569
	(0.368)	(0.148)	(0.368)
	p=0.648	*p=0.892*	*p=0.129*
Difference	0.678*	0.364**	0.310
	(0.368)	(0.155)	(0.368)
	p=0.065	*p=0.021*	*p=0.371*
N	544	544	544

Note. This table reproduces Table 6 for teachers who were employed in the HI at baseline. Single-hypothesis two-sided p-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations). 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	0.215	-3.280	-7.272	2.034
	(0.287)	(13.123)	(7.774)	(10.258)
	p=0.456	*p=0.803*	*p=0.359*	*p=0.831*
FE+HIM	0.252	1.521	6.599	-6.345
	(0.291)	(13.953)	(9.944)	(9.269)
	p=0.391	*p=0.904*	*p=0.517*	*p=0.487*
Difference	0.038	4.801	13.872	-8.379
	(0.238)	(13.953)	(9.949)	(9.291)
	p=0.874	*p=0.747*	*p=0.163*	*p=0.369*
N	1075	1065	836	1075

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. Single-hypothesis two-sided p-values calculated using block bootstraps, resampling triplets with replacement (1,000 iterations). 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 Anthropometic Outcomes

	(1) Weight for Age Z-Score	(2) Length for Age Z-Score	(3) BMI for Age Z-Score	(4) Weight for Length Z-Score	(5) Acute Malnutrition	(6) Obese	(7) Chronic Malnutrition
HIM	0.033	0.059	-0.010	-0.003	-0.003	0.021	-0.032**
	(0.052)	(0.038)	(0.070)	(0.071)	(0.014)	(0.016)	(0.016)
	p=0.548	*p=0.121*	*p=0.890*	*p=0.968*	*p=0.832*	*p=0.168*	*p=0.041*
	q=0.862	*q=0.419*	*q=0.907*	*q=0.968*	*q=0.976*	*q=0.441*	*q=0.200*
HIM+FE	0.082	0.013	0.105	0.117	-0.001	0.027*	-0.028
	(0.058)	(0.034)	(0.077)	(0.079)	(0.013)	(0.015)	(0.018)
	p=0.166	*p=0.690*	*p=0.170*	*p=0.130*	*p=0.909*	*p=0.080*	*p=0.101*
	q=0.442	*q=0.901*	*q=0.438*	*q=0.363*	*q=0.909*	*q=0.328*	*q=0.346*
Difference	0.049	-0.046	0.115*	0.120*	0.002	0.006	0.004
	(0.054)	(0.034)	(0.070)	(0.071)	(0.013)	(0.015)	(0.016)
	p=0.365	*p=0.172*	*p=0.097*	*p=0.083*	*p=0.900*	*p=0.705*	*p=0.829*
	q=0.819	*q=0.578*	*q=0.382*	*q=0.337*	*q=0.900*	*q=0.971*	*q=0.971*
N	1066	1065	1065	1065	1065	1065	1065
Control mean	-0.446	-0.813	0.104	0.071	0.023	0.043	0.123
Control SD	1.058	1.014	1.093	1.114	0.150	0.203	0.329

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. Single-hypothesis two-sided p-values calculated using block bootstraps, 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. (2016). Standard errors (bootstrapped) in parentheses. All estimates control for age, gender, city effects, and baseline values of the outcome variables.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C Measurement System Parameters

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

Assessment	**Item (j)**	β_j		α_j			g_j	
Daberon	1	1.10	[.57, 1.71]	α_{j1}	4.73	[3.78, 5.78]	0.01	[0, .37]
Daberon	2	1.72	[1.16, 2.64]	α_{j1}	2.25	[1.71, 2.95]	0.33	[.01, .5]
Daberon	3	1.35	[1.03, 2.01]	α_{j1}	1.43	[.7, 1.9]	0.21	[0, .48]
Daberon	4	0.66	[.15, .9]	α_{j1}	3.92	[3.43, 4.34]	0.03	[0, .15]
Daberon	5	0.78	[.61, 2.56]	α_{j1}	2.54	[.12, 2.85]	0.01	[0, .81]
Daberon	6	0.80	[.69, 4.22]	α_{j1}	1.71	[-.73, 1.93]	0.02	[0, .7]
Daberon	7	0.76	[.33, 1.13]	α_{j1}	3.27	[2.78, 3.65]	0.01	[0, .24]
Daberon	8	1.10	[.74, 2.5]	α_{j1}	4.20	[2.63, 4.75]	0.01	[0, .85]
Daberon	9	0.88	[.62, 5.54]	α_{j1}	4.03	[1.9, 4.56]	0.02	[0, .92]
Daberon	10	1.46	[1.22, 1.97]	α_{j1}	1.34	[.92, 1.72]	0.07	[0, .22]
Daberon	11	1.57	[1.19, 2.19]	α_{j1}	2.00	[1.53, 2.45]	0.20	[0, .4]
Daberon	12	5.34	[.59, 6.36]	α_{j1}	2.53	[1.96, 5.28]	0.94	[.01, .94]
Daberon	14	3.28	[2.53, 6.7]	α_{j1}	1.91	[1.34, 2.74]	0.23	[.08, .43]
Daberon	15	3.87	[2.91, 8.66]	α_{j1}	2.28	[1.7, 3.36]	0.18	[.02, .38]
Daberon	16	3.27	[2.38, 5.35]	α_{j1}	1.56	[.99, 2.27]	0.27	[.15, .36]
Daberon	17	2.18	[1.66, 3.59]	α_{j1}	1.33	[.84, 1.75]	0.08	[0, .26]
Daberon	18	3.24	[2.18, 5.67]	α_{j1}	1.68	[1.06, 2.4]	0.39	[.19, .53]
Daberon	19	1.98	[1.39, 3.3]	α_{j1}	0.97	[.41, 1.47]	0.27	[.01, .44]
Daberon	20	1.85	[1.46, 2.47]	α_{j1}	1.19	[.89, 1.51]	0.00	[0, 0]
Daberon	21	0.98	[.71, 2.43]	α_{j1}	0.73	[-.51, 1.15]	0.14	[0, .44]
Daberon	22	1.38	[1.06, 1.96]	α_{j1}	-0.35	[-.89, -.03]	0.07	[0, .19]
Daberon	23	0.52	[-.03, .98]	α_{j1}	3.67	[3.13, 4.2]	0.06	[.01, .19]
Daberon	24	1.61	[1.26, 2.52]	α_{j1}	0.36	[-.17, .64]	0.05	[0, .21]
Daberon	25	1.33	[.93, 3.09]	α_{j1}	1.34	[.17, 2.22]	0.44	[.01, .7]
Daberon	26	1.30	[1.01, 1.7]	α_{j1}	-0.24	[-.52, -.01]	0.00	[0, 0]
Daberon	27	3.16	[1.84, 5.3]	α_{j1}	1.67	[.81, 2.69]	0.55	[.26, .66]
Daberon	28	1.68	[1.29, 2.16]	α_{j1}	0.50	[.1, .85]	0.13	[0, .22]
Daberon	29	1.69	[1.25, 2.34]	α_{j1}	-0.74	[-1.28, -.28]	0.15	[.05, .22]
Daberon	30	1.66	[1.16, 4.03]	α_{j1}	1.63	[.41, 2.43]	0.49	[.02, .74]
Daberon	31	0.83	[.4, 1.34]	α_{j1}	1.91	[1.23, 3.8]	0.76	[.02, .85]
Daberon	32	2.53	[1.04, 8.55]	α_{j1}	-2.85	[-10.17, -.91]	0.41	[.25, .47]
Daberon	33	1.48	[1.11, 2.43]	α_{j1}	1.75	[1.04, 2.24]	0.24	[0, .54]
Daberon	34	1.67	[1.1, 2.54]	α_{j1}	-0.64	[-1.34, -.08]	0.17	[.02, .27]
Daberon	35	1.99	[1.19, 4.91]	α_{j1}	-0.37	[-2.46, .59]	0.57	[.43, .66]
Daberon	36	1.27	[.57, 3.07]	α_{j1}	-0.52	[-1.94, .82]	0.50	[.02, .61]
Daberon	37	1.04	[.67, 1.81]	α_{j1}	1.02	[.13, 1.81]	0.45	[.02, .65]
Daberon	38	0.74	[.57, 1.03]	α_{j1}	-0.38	[-.62, -.24]	0.00	[0, .03]
Daberon	39	1.07	[.72, 2.1]	α_{j1}	-1.01	[-2.36, -.3]	0.25	[.1, .37]
Daberon	41	0.53	[.11, 3.01]	α_{j1}	4.22	[2.46, 4.47]	0.21	[.15, .92]
Daberon	42	2.16	[1.14, 5.38]	α_{j1}	-0.72	[-3.56, .73]	0.77	[.62, .82]
Daberon	43	1.06	[.87, 1.55]	α_{j1}	0.75	[.31, .97]	0.00	[0, .19]
Daberon	44	2.38	[1.54, 4.16]	α_{j1}	-1.56	[-3.33, -.64]	0.14	[0, .21]
Daberon	45	0.65	[.32, 1.12]	α_{j1}	2.54	[1.34, 3.18]	0.29	[.01, .74]
Daberon	46	1.02	[.61, 1.55]	α_{j1}	1.70	[1.05, 2.47]	0.44	[.02, .66]
Daberon	47	0.84	[.63, 1.2]	α_{j1}	2.85	[1.92, 3.24]	0.02	[0, .51]
Daberon	48	0.32	[.08, .86]	α_{j1}	1.93	[.71, 2.22]	0.05	[.02, .66]
Daberon	49	1.68	[1.36, 2.1]	α_{j1}	1.94	[1.58, 2.35]	0.00	[0, 0]
Daberon	50	0.41	[.31, 1.08]	α_{j1}	0.44	[-.99, .56]	0.01	[0, .43]
Daberon	51	0.87	[.63, 1.19]	α_{j1}	1.99	[1.74, 2.32]	0.00	[0, .02]
Daberon	52	0.79	[.62, 1.34]	α_{j1}	1.62	[.7, 1.85]	0.00	[0, .44]
Daberon	53	0.77	[.6, .97]	α_{j1}	1.63	[1.4, 1.85]	0.00	[0, .02]
Daberon	54	1.02	[.79, 1.28]	α_{j1}	1.60	[1.36, 1.87]	0.00	[0, 0]
Daberon	55	1.00	[.81, 1.22]	α_{j1}	2.20	[1.87, 2.58]	0.00	[0, 0]
Daberon	56	1.33	[.92, 2.22]	α_{j1}	0.29	[-.62, 1.02]	0.29	[0, .45]
Daberon	57	0.74	[.6, 1.91]	α_{j1}	-0.60	[-2.06, -.47]	0.00	[0, .22]
Daberon	58	0.03	[-.48, 1.04]	α_{j1}	0.03	[-3.66, .52]	0.20	[.01, .57]
Daberon	59	0.52	[.42, 1.13]	α_{j1}	0.78	[-.38, .96]	0.01	[0, .42]
Daberon	60	0.72	[.49, 1.16]	α_{j1}	-1.06	[-1.46, -.86]	0.03	[0, .12]
Daberon	61	0.23	[-.14, .58]	α_{j1}	2.17	[1.64, 2.66]	0.07	[.01, .23]
Daberon	62	0.87	[.61, 2.56]	α_{j1}	1.05	[-1.02, 1.67]	0.39	[.02, .72]
Daberon	63	1.19	[.96, 1.55]	α_{j1}	0.06	[-.27, .27]	0.00	[0, 0]
Daberon	64	1.08	[.84, 2.25]	α_{j1}	-1.52	[-3, -1.26]	0.02	[0, .13]
Daberon	65	1.56	[1.17, 2.27]	α_{j1}	-0.23	[-.73, .13]	0.00	[0, .06]
Daberon	66	1.97	[1.56, 2.72]	α_{j1}	-0.06	[-.62, .32]	0.00	[0, .07]
Daberon	67	1.36	[1.07, 2.02]	α_{j1}	-0.32	[-.84, -.02]	0.00	[0, .07]
Daberon	68	1.53	[.96, 2.97]	α_{j1}	-1.85	[-3.17, -1.4]	0.03	[0, .1]
Daberon	69	1.58	[.98, 2.92]	α_{j1}	-1.51	[-2.6, -1.08]	0.05	[0, .12]

Daberon	70	0.90	[.64, 1.86]	α_{j1}	-0.97	[-1.96, -.8]	0.00	[0, .12]
PTT		0.89	[.73, 1.04]	α_{j1}	3.20	[2.94, 3.55]		
PTT				α_{j2}	2.43	[2.2, 2.7]		
PTT				α_{j3}	1.90	[1.69, 2.12]		
PTT				α_{j4}	1.40	[1.19, 1.6]		
PTT				α_{j5}	0.91	[.71, 1.11]		
PTT				α_{j6}	0.59	[.4, .78]		
PTT				α_{j7}	0.32	[.14, .5]		
PTT				α_{j8}	-0.05	[-.25, .14]		
PTT				α_{j9}	-0.43	[-.62, -.25]		
PTT				α_{j10}	-0.77	[-.98, -.56]		
PTT				α_{j11}	-1.26	[-1.47, -1.06]		
PTT				α_{j12}	-1.67	[-1.9, -1.46]		
PTT				α_{j13}	-1.99	[-2.24, -1.76]		
PTT				α_{j14}	-2.31	[-2.62, -2.05]		
PTT				α_{j15}	-2.84	[-3.24, -2.54]		
PTT				α_{j16}	-3.41	[-3.88, -3.04]		
TVIP	1	4.57	[2.72, 17.34]	α_{j1}	5.41	[4.44, 14.53]	0.77	[.52, .89]
TVIP	2	2.17	[1.33, 4.04]	α_{j1}	3.00	[2.18, 4.23]	0.46	[.01, .7]
TVIP	3	1.64	[1.14, 3.41]	α_{j1}	2.18	[1.06, 3.03]	0.52	[.07, .8]
TVIP	4	1.16	[.83, 1.72]	α_{j1}	1.88	[1.15, 2.66]	0.32	[.01, .54]
TVIP	5	1.77	[1.08, 2.62]	α_{j1}	2.69	[1.95, 3.47]	0.51	[.24, .71]
TVIP	6	0.66	[.48, 1.03]	α_{j1}	1.69	[.69, 1.91]	0.00	[0, .51]
TVIP	7	1.83	[1.3, 2.75]	α_{j1}	1.26	[.67, 1.8]	0.28	[.12, .45]
TVIP	8	1.02	[.51, 2.1]	α_{j1}	-0.03	[-1.19, 1.25]	0.54	[.02, .68]
TVIP	9	1.00	[.66, 1.56]	α_{j1}	1.31	[.66, 1.85]	0.26	[0, .5]
TVIP	10	0.93	[.66, 1.74]	α_{j1}	0.93	[.03, 1.42]	0.23	[0, .51]
TVIP	11	0.88	[.48, 1.76]	α_{j1}	2.53	[1.84, 3.02]	0.20	[0, .63]
TVIP	12	1.53	[.86, 4.76]	α_{j1}	1.11	[-.41, 2.13]	0.53	[.01, .75]
TVIP	13	2.04	[.9, 11.61]	α_{j1}	-1.08	[-10.06, .39]	0.68	[.49, .75]
TVIP	14	1.45	[1, 2.74]	α_{j1}	0.38	[-.78, 1.27]	0.42	[.01, .6]
TVIP	15	1.69	[.92, 4.24]	α_{j1}	0.08	[-1.92, 1.32]	0.53	[.02, .68]
TVIP	16	2.10	[1.4, 4.32]	α_{j1}	1.86	[.82, 2.66]	0.42	[0, .69]
TVIP	17	3.16	[2.13, 6.36]	α_{j1}	0.60	[-.62, 1.39]	0.42	[.24, .56]
TVIP	18	2.63	[1.83, 5.08]	α_{j1}	0.72	[-.41, 1.46]	0.43	[.22, .58]
TVIP	19	1.10	[.74, 2.64]	α_{j1}	0.55	[-1.16, 1.16]	0.27	[0, .57]
TVIP	20	1.35	[.88, 2.18]	α_{j1}	1.25	[.63, 1.86]	0.37	[.02, .57]
TVIP	21	1.52	[1, 2.7]	α_{j1}	0.04	[-.82, .72]	0.32	[0, .49]
TVIP	22	1.13	[.7, 1.82]	α_{j1}	-0.24	[-1.03, .52]	0.39	[.12, .5]
TVIP	23	1.31	[1.08, 3.06]	α_{j1}	1.28	[-.54, 1.74]	0.17	[0, .63]
TVIP	24	2.36	[1.2, 5.63]	α_{j1}	-1.47	[-4.31, -.03]	0.51	[.26, .6]
TVIP	25	1.27	[1.01, 5.57]	α_{j1}	1.64	[-1.11, 1.89]	0.00	[0, .67]
TVIP	26	2.26	[1.57, 3.83]	α_{j1}	0.19	[-.71, .86]	0.61	[.48, .7]
TVIP	27	2.29	[1.49, 4.64]	α_{j1}	-0.93	[-2.59, -.04]	0.55	[.4, .65]
TVIP	28	0.90	[.64, 4.2]	α_{j1}	0.43	[-3.46, 1.08]	0.30	[0, .67]
TVIP	29	0.79	[.53, 1.57]	α_{j1}	0.39	[-.58, .59]	0.00	[0, .33]
TVIP	30	0.96	[.77, 2.53]	α_{j1}	0.66	[-1.4, .84]	0.01	[0, .52]
TVIP	31	1.34	[1.07, 2.88]	α_{j1}	-0.50	[-2.22, -.27]	0.02	[0, .26]
TVIP	32	0.74	[.15, 5.92]	α_{j1}	-0.34	[-6.51, .93]	0.49	[.01, .67]
TVIP	33	1.03	[.81, 4.24]	α_{j1}	0.12	[-3.62, .29]	0.01	[0, .47]
TVIP	34	0.50	[.15, 5.6]	α_{j1}	2.81	[-1.69, 3.69]	0.25	[.14, .92]
TVIP	35	3.36	[1.51, 9]	α_{j1}	-1.32	[-5.22, .63]	0.52	[.04, .61]
TVIP	36	0.65	[.14, 1.7]	α_{j1}	-0.72	[-2.18, .68]	0.51	[.04, .63]
TVIP	37	1.27	[.85, 2.84]	α_{j1}	-1.05	[-3.26, -.25]	0.21	[0, .39]
TVIP	38	0.66	[-.01, 4.75]	α_{j1}	-1.55	[-7.91, -.33]	0.32	[.01, .45]
TVIP	39	1.71	[1.19, 3.81]	α_{j1}	-1.38	[-3.69, -.59]	0.17	[0, .31]
TVIP	40	2.62	[1.86, 5.22]	α_{j1}	-3.00	[-5.45, -1.96]	0.33	[.24, .4]
TVIP	41	1.24	[.41, 3.52]	α_{j1}	-1.81	[-4.87, .02]	0.45	[.05, .55]
TVIP	42	2.01	[-.02, 4.59]	α_{j1}	-3.82	[-8.97, -1.93]	0.22	[.09, .29]
TVIP	43	1.32	[1.08, 12.42]	α_{j1}	0.30	[-9.92, .83]	0.20	[0, .64]
WM-12 (a)		1.06	[.87, 1.26]	α_{j1}	3.04	[2.73, 3.43]		
WM-12 (a)				α_{j2}	2.18	[1.91, 2.49]		
WM-12 (a)				α_{j3}	1.39	[1.13, 1.69]		
WM-12 (a)				α_{j4}	0.74	[.52, .97]		
WM-12 (a)				α_{j5}	0.10	[-.12, .32]		
WM-12 (a)				α_{j6}	-0.63	[-.84, -.4]		
WM-12 (a)				α_{j7}	-1.31	[-1.51, -1.1]		
WM-12 (a)				α_{j8}	-1.98	[-2.21, -1.77]		
WM-12 (a)				α_{j9}	-2.45	[-2.71, -2.2]		
WM-12 (a)				α_{j10}	-3.23	[-3.54, -2.95]		
WM-12 (a)				α_{j11}	-4.11	[-4.57, -3.77]		

WM-12 (c)		1.37	[1.13, 1.68]	α_{j1}	2.98	[2.71, 3.37]		
WM-12 (c)				α_{j2}	2.07	[1.79, 2.41]		
WM-12 (c)				α_{j3}	1.19	[.91, 1.5]		
WM-12 (c)				α_{j4}	0.52	[.25, .78]		
WM-12 (c)				α_{j5}	-0.19	[-.47, .09]		
WM-12 (c)				α_{j6}	-0.84	[-1.09, -.6]		
WM-12 (c)				α_{j7}	-1.83	[-2.11, -1.56]		
WM-12 (c)				α_{j8}	-2.43	[-2.72, -2.14]		
WM-12 (c)				α_{j9}	-2.89	[-3.25, -2.57]		
WM-12 (c)				α_{j10}	-3.50	[-3.93, -3.12]		
WM-12 (c)				α_{j11}	-4.08	[-4.61, -3.64]		
WM-14	1	1.43	[.86, 4.81]	α_{j1}	3.34	[2.22, 5.28]	0.72	[.05, .91]
WM-14	4	0.15	[-.07, .42]	α_{j1}	1.37	[.87, 1.69]	0.08	[.01, .24]
WM-14	5	1.41	[.89, 4.72]	α_{j1}	5.88	[4.29, 13.61]	0.07	[0, .75]
WM-14	6	0.94	[.41, 1.61]	α_{j1}	4.01	[3.36, 5.18]	0.01	[0, .14]
WM-14	7	2.76	[1.81, 5.74]	α_{j1}	4.64	[4.01, 7.07]	0.30	[0, .53]
WM-14	8	1.11	[.74, 1.62]	α_{j1}	3.37	[2.77, 4.05]	0.00	[0, .27]
WM-14	9	0.83	[.64, 2.5]	α_{j1}	3.58	[1.18, 4.24]	0.05	[.01, .9]
WM-14	10	1.01	[.79, 1.37]	α_{j1}	2.32	[1.81, 2.65]	0.00	[0, .31]
WM-14	11	1.73	[1.38, 2.51]	α_{j1}	4.21	[3.5, 5.12]	0.01	[0, .4]
WM-14	12	1.28	[.86, 1.74]	α_{j1}	3.90	[3.32, 4.5]	0.00	[0, .23]
WM-14	13	2.71	[2, 4.13]	α_{j1}	3.90	[3.25, 4.9]	0.19	[0, .49]
WM-14	14	1.40	[1.09, 2.05]	α_{j1}	2.32	[1.92, 2.73]	0.03	[0, .28]
WM-14	15	1.00	[.79, 1.55]	α_{j1}	1.46	[.8, 1.85]	0.16	[0, .44]
WM-14	16	1.10	[.94, 1.66]	α_{j1}	0.75	[-.01, 1]	0.04	[0, .35]
WM-14	17	1.20	[.93, 1.59]	α_{j1}	0.73	[.44, .98]	0.00	[0, .01]
WM-14	18	1.72	[1.39, 2.44]	α_{j1}	1.83	[1.34, 2.16]	0.03	[0, .26]
WM-14	19	1.91	[1.56, 2.75]	α_{j1}	0.36	[-.13, .67]	0.00	[0, .09]
WM-14	20	1.51	[1.21, 1.97]	α_{j1}	0.33	[-.01, .62]	0.00	[0, 0]
WM-14	21	1.98	[1.71, 2.47]	α_{j1}	-2.30	[-2.75, -1.97]	0.00	[0, 0]
WM-14	22	1.47	[1.2, 1.84]	α_{j1}	-1.29	[-1.64, -1]	0.00	[0, 0]
WM-14	23	1.64	[1.32, 2.23]	α_{j1}	-1.66	[-2.09, -1.36]	0.00	[0, 0]
WM-14	24	1.48	[1.17, 2.04]	α_{j1}	0.11	[-.21, .38]	0.00	[0, 0]
WM-14	25	0.96	[.48, 2.04]	α_{j1}	-3.90	[-5.19, -3.34]	0.00	[0, 0]
WM-14	26	1.21	[.84, 1.78]	α_{j1}	-4.21	[-5.06, -3.63]	0.00	[0, 0]
WM-14	27	0.17	[-.55, .88]	α_{j1}	-4.16	[-4.75, -3.77]	0.00	[0, 0]
WM-17	1	0.49	[-.47, 1.04]	α_{j1}	4.33	[3.63, 5.04]	0.09	[.01, .16]
WM-17	4	0.76	[.49, 1.26]	α_{j1}	2.15	[1.31, 2.61]	0.11	[0, .58]
WM-17	5	0.42	[0, 1.04]	α_{j1}	2.70	[1.89, 3.23]	0.10	[.02, .68]
WM-17	6	0.88	[.25, 2.32]	α_{j1}	0.69	[-.73, 2.26]	0.74	[.15, .86]
WM-17	7	1.33	[1.02, 1.88]	α_{j1}	1.88	[1.51, 2.24]	0.00	[0, .13]
WM-17	8	1.05	[.87, 1.46]	α_{j1}	1.23	[.78, 1.48]	0.00	[0, .22]
WM-17	9	1.56	[1.03, 2.43]	α_{j1}	0.49	[-.14, 1.01]	0.23	[0, .36]
WM-17	10	0.78	[.56, 1.62]	α_{j1}	1.01	[-.26, 1.23]	0.01	[0, .45]
WM-17	11	1.02	[.85, 3.21]	α_{j1}	1.59	[-.95, 1.81]	0.02	[0, .69]
WM-17	12	0.95	[.75, 1.36]	α_{j1}	1.26	[.72, 1.49]	0.00	[0, .29]
WM-17	13	1.92	[.52, 6.14]	α_{j1}	-2.31	[-7.26, .15]	0.42	[0, .47]
WM-17	14	1.39	[.69, 3.42]	α_{j1}	-2.25	[-4.57, -1.29]	0.10	[0, .17]
WM-17	15	0.84	[.7, 4.35]	α_{j1}	-1.17	[-5.85, -.99]	0.01	[0, .24]

Notes: Table presents estimated parameters for IRT measurement model of children's cognitive development alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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	[.93, 1.58]	-1.64	[-2.01, -1.57]	-4.60	[-5.22, -4.35]
2	0.63	[.42, 1.38]	-3.36	[-4.26, -3.16]	-5.36	[-6.22, -5.08]
3	0.44	[.25, .87]	-0.90	[-1.35, -.74]	-2.64	[-3.82, -2.08]
4	0.06	[-.03, .22]	2.76	[2.45, 3.3]	0.69	[.55, .84]
5	1.14	[.79, 1.62]	-0.55	[-.89, -.38]	-2.49	[-3.28, -2.09]
6	-0.03	[-.29, .05]	-0.06	[-.18, .15]	-1.41	[-1.53, -1.21]
7	1.30	[1.08, 1.7]	-0.81	[-1.22, -.63]	-3.97	[-4.73, -3.37]
8	0.91	[.7, 1.15]	-0.87	[-1.21, -.66]	-3.40	[-3.92, -3.21]
9	1.34	[1.12, 2.19]	-3.30	[-4.52, -3.08]	-5.47	[-7.67, -4.85]
10	0.95	[.9, 1.32]	-1.79	[-2.3, -1.68]	-4.19	[-5.02, -3.78]
11	1.32	[.99, 1.56]	0.33	[.14, .4]	-4.30	[-5.29, -4.02]
12	-0.16	[-.41, -.03]	2.94	[2.71, 3.35]	0.42	[.24, .67]
13	0.49	[.23, .61]	-1.07	[-1.31, -.95]	-3.37	[-3.93, -3.02]
14	0.80	[.38, 1]	-2.55	[-3, -2.1]	-4.36	[-5.33, -3.71]
15	0.71	[.45, 1.64]	-2.70	[-3.75, -2.22]	-3.78	[-5.29, -3.1]
16	1.82	[1.42, 2.72]	-4.26	[-5.82, -3.69]	-7.40	[-8.42, -6.17]
17	0.87	[.59, 1.11]	0.47	[.32, .64]	-1.98	[-2.31, -1.8]
18	0.56	[.41, .88]	-0.79	[-1.03, -.39]	-2.15	[-2.37, -1.79]
19	0.98	[.56, 1.68]	-3.53	[-4.37, -3.39]	-5.62	[-6.79, -4.77]
20	1.04	[.72, 1.22]	-2.04	[-2.2, -1.69]	-4.28	[-4.73, -3.68]
21	0.76	[.57, .98]	-1.19	[-1.4, -.96]	-3.53	[-4.41, -3.14]
22	1.43	[1.28, 1.89]	-2.73	[-3.19, -2.55]	-5.71	[-6.93, -5.22]
23	1.07	[.62, 1.48]	-2.98	[-3.64, -2.62]	-5.13	[-5.95, -4.39]
24	1.30	[.96, 1.82]	-2.68	[-3.35, -2.44]	-5.22	[-6.15, -4.8]
25	1.37	[.7, 2.27]	-2.74	[-3.54, -2.18]	-5.32	[-6.77, -4.3]
26	0.79	[.57, .97]	-1.94	[-2.39, -1.6]	-4.54	[-5.35, -4.02]
27	0.11	[-.1, .38]	-0.14	[-.45, .07]	-2.09	[-2.44, -1.79]
28	0.17	[-.17, .57]	-2.07	[-2.41, -1.93]	-2.93	[-3.52, -2.69]
29	1.65	[1.25, 2.38]	-2.95	[-4.16, -2.49]	-6.42	[-7.86, -6.01]
30	0.62	[.55, .83]	-0.50	[-.8, -.34]	-2.37	[-2.72, -2.21]
31	1.23	[.95, 1.39]	-0.41	[-.64, -.06]	-4.00	[-4.63, -3.45]
32	0.85	[.68, 1.15]	-1.44	[-1.8, -1.11]	-3.23	[-4.33, -2.82]

Notes: Table presents estimated parameters for IRT measurement model of children's socioemotional problems alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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

Panel A: ASQ-Communication				
Item (j)	β_j		α_j	
Item 1	0.70	(0.28)	1.84	(0.31)
Item 2	1.94	(0.50)	1.00	(0.36)
Item 3	0.20	(0.23)	2.44	(0.20)
Item 4	1.98	(0.39)	3.11	(0.43)
Item 5	2.22	(0.39)	4.03	(0.47)
Item 6	0.66	(0.27)	0.54	(0.17)
Item 7	0.52	(0.29)	0.35	(0.20)
Item 8	0.55	(0.21)	1.01	(0.15)
Item 9	1.86	(0.87)	3.39	(0.82)
Item 10	0.84	(0.55)	2.82	(0.46)
Item 11	1.09	(0.46)	1.36	(0.29)
Item 12	1.25	(0.24)	2.04	(0.19)
Item 13	0.47	(0.30)	0.69	(0.19)
Item 14	1.72	(0.66)	0.21	(0.35)
Item 15	1.91	(0.43)	3.06	(0.50)
Item 16	3.43	(1.44)	0.98	(0.38)
Item 17	1.62	(0.43)	1.88	(0.43)
Item 18	0.37	(0.25)	1.99	(0.19)
Item 19	1.09	(0.24)	0.44	(0.12)
Item 20	1.38	(0.23)	1.26	(0.14)
Item 21	1.05	(0.20)	1.14	(0.13)

Panel B: ASQ-Gross Motor				
Item (j)	β_j		α_j	
Item 1	0.49	(0.63)	2.81	(0.75)
Item 2	0.91	(0.18)	1.77	(0.15)
Item 3	1.99	(0.30)	2.05	(0.23)
Item 4	1.91	(0.35)	3.51	(0.39)
Item 5	0.76	(0.19)	1.40	(0.16)
Item 6	1.52	(0.44)	4.05	(0.56)
Item 7	1.40	(0.40)	2.58	(0.33)
Item 8	1.28	(0.20)	0.40	(0.11)
Item 9	0.54	(0.32)	0.17	(0.26)
Item 10	0.82	(0.17)	1.19	(0.12)
Item 11	1.25	(0.21)	0.76	(0.12)
Item 12	1.35	(0.54)	1.28	(0.31)
Item 13	2.42	(1.32)	4.25	(1.73)
Item 14	0.85	(0.39)	0.15	(0.24)
Item 15	0.96	(0.33)	1.17	(0.21)
Item 16	2.14	(0.49)	0.62	(0.25)

Panel C: ASQ-Fine Motor				
Item (j)	β_j		α_j	
Item 1	1.15	(0.40)	0.61	(0.26)
Item 2	0.53	(0.50)	0.35	(0.36)
Item 3	1.27	(0.34)	3.41	(0.40)
Item 4	-0.03	(0.41)	2.54	(0.35)
Item 5	1.28	(0.27)	-1.60	(0.22)
Item 6	0.56	(0.14)	0.72	(0.11)
Item 7	0.49	(0.24)	-0.04	(0.19)
Item 8	1.04	(0.38)	-1.93	(0.34)
Item 9	3.47	(1.37)	0.82	(0.46)
Item 10	1.11	(0.25)	-0.25	(0.18)
Item 11	0.30	(0.16)	1.54	(0.13)
Item 12	0.81	(0.22)	0.53	(0.15)
Item 13	1.58	(0.24)	0.94	(0.16)
Item 14	0.49	(0.18)	-1.07	(0.15)
Item 15	4.34	(1.06)	0.43	(0.29)
Item 16	4.09	(1.40)	-0.60	(0.43)
Item 17	4.31	(1.17)	0.65	(0.34)
Item 18	0.77	(0.34)	-1.40	(0.31)
Item 19	1.09	(0.62)	3.29	(0.72)
Item 20	0.48	(0.13)	0.35	(0.11)
Item 21	8.84	(9.30)	-3.00	(3.13)

Panel D: ASQ-Problem Solving				
Item (j)	β_j		α_j	
Item 1	0.65	(0.83)	3.17	(0.83)
Item 2	1.25	(0.81)	-0.15	(0.31)
Item 3	0.58	(0.36)	3.51	(0.35)
Item 4	0.07	(0.19)	2.37	(0.14)
Item 5	0.12	(0.18)	2.02	(0.14)
Item 6	0.41	(0.23)	0.97	(0.18)
Item 7	0.63	(0.30)	0.31	(0.20)
Item 8	0.97	(0.32)	0.45	(0.20)
Item 9	0.82	(0.27)	1.60	(0.22)
Item 10	0.69	(0.20)	1.03	(0.14)
Item 11	1.18	(0.34)	0.44	(0.20)
Item 12	0.97	(0.17)	0.65	(0.11)
Item 13	0.44	(0.24)	1.76	(0.20)
Item 14	2.21	(1.09)	2.08	(0.77)
Item 15	2.20	(0.40)	0.74	(0.19)
Item 16	1.07	(0.54)	0.29	(0.32)
Item 17	0.99	(0.41)	1.49	(0.30)
Item 18	1.32	(0.71)	1.81	(0.60)
Item 19	3.97	(1.39)	-1.35	(0.48)
Item 20	1.27	(0.24)	1.89	(0.19)
Item 21	0.99	(0.21)	2.08	(0.17)
Item 22	0.92	(0.37)	-1.12	(0.31)

Panel E: ASQ-Socio Individual				
Item (j)	β_j		α_j	
Item 1	0.45	(0.55)	3.19	(0.53)
Item 2	0.92	(0.27)	1.37	(0.21)
Item 3	0.47	(0.20)	-0.74	(0.14)
Item 4	0.67	(0.18)	-0.13	(0.11)
Item 5	1.04	(0.31)	1.42	(0.19)
Item 6	1.26	(0.22)	-0.07	(0.10)
Item 7	1.08	(0.19)	1.15	(0.12)
Item 8	0.55	(0.21)	1.62	(0.15)
Item 9	1.88	(0.54)	0.64	(0.24)
Item 10	0.60	(0.17)	0.10	(0.11)
Item 11	0.87	(0.30)	2.30	(0.25)
Item 12	0.39	(0.20)	2.09	(0.15)
Item 13	1.67	(0.33)	2.69	(0.30)
Item 14	1.26	(0.61)	0.86	(0.27)

Panel F: ASQ-Socio Emotional				
Item (j)	β_j		α_j	
Item 1	-0.38	(0.40)	2.28	(0.33)
Item 2	1.55	(1.06)	-5.33	(1.61)
Item 3	0.07	(0.34)	-1.84	(0.26)
Item 4	0.46	(0.51)	-2.99	(0.45)
Item 5	1.07	(0.25)	-3.54	(0.29)
Item 6	0.64	(0.30)	-1.00	(0.22)
Item 7	-0.25	(0.38)	2.08	(0.29)
Item 8	-0.06	(0.11)	1.18	(0.09)
Item 9	1.64	(0.25)	-2.75	(0.24)
Item 10	-0.24	(0.31)	1.57	(0.24)
Item 11	-0.22	(0.22)	3.07	(0.19)
Item 12	-0.08	(0.49)	2.74	(0.37)
Item 13	0.65	(0.13)	-1.47	(0.11)
Item 14	2.87	(1.41)	-4.63	(1.72)
Item 15	1.32	(0.24)	-3.10	(0.26)
Item 16	-0.81	(0.58)	3.36	(0.61)
Item 17	1.20	(0.23)	-2.95	(0.24)
Item 18	1.17	(0.44)	-2.04	(0.39)
Item 19	0.85	(0.43)	-2.42	(0.40)
Item 20	0.98	(0.19)	-2.70	(0.20)
Item 21	1.16	(0.19)	-2.49	(0.19)
Item 22	0.46	(0.44)	4.38	(0.39)
Item 23	0.90	(0.44)	-2.37	(0.41)
Item 24	0.28	(0.21)	2.79	(0.18)
Item 25	0.48	(0.12)	-0.55	(0.10)
Item 26	0.86	(0.17)	-1.92	(0.15)
Item 27	0.54	(0.13)	-1.11	(0.11)
Item 28	1.23	(0.32)	-4.06	(0.41)
Item 29	1.06	(0.16)	0.22	(0.11)
Item 30	0.72	(0.13)	0.06	(0.10)
Item 31	1.27	(0.18)	-0.88	(0.13)
Item 32	0.71	(0.14)	0.85	(0.11)
Item 33	1.42	(0.25)	-2.78	(0.26)
Item 34	1.44	(0.20)	-1.04	(0.14)
Item 35	1.05	(0.16)	-0.93	(0.12)
Item 36	1.59	(0.22)	-1.12	(0.15)
Item 37	0.03	(0.17)	2.16	(0.14)
Item 38	1.59	(0.33)	-3.81	(0.40)
Item 39	1.41	(0.23)	-2.43	(0.22)
Item 40	0.26	(0.14)	1.43	(0.12)

Panel G: MacArthur-Bates - Younger Kids

Item (j)	Word	β_j		α_j	
Item 1	A	0.63	(0.21)	2.84	(0.25)
Item 2	Adiós/chao	1.77	(0.33)	4.39	(0.53)
Item 3	Afuera	2.99	(0.36)	1.19	(0.23)
Item 4	Aquí	1.91	(0.25)	2.14	(0.24)
Item 5	Arroz	2.19	(0.31)	3.09	(0.33)
Item 6	Baño	4.33	(0.64)	3.75	(0.53)
Item 7	Besar	1.89	(0.23)	0.98	(0.17)
Item 8	Bigote	2.73	(0.38)	-2.51	(0.32)
Item 9	Bonita/linda	2.36	(0.30)	1.90	(0.24)
Item 10	Brazo	2.25	(0.27)	0.73	(0.18)
Item 11	Buenas noches	2.14	(0.26)	-0.09	(0.17)
Item 12	Bus	1.89	(0.25)	2.01	(0.22)
Item 13	Caer(se)	2.64	(0.34)	2.33	(0.28)
Item 14	Caliente	2.78	(0.38)	3.22	(0.38)
Item 15	Calle	2.10	(0.26)	1.71	(0.21)
Item 16	Cama	4.37	(0.69)	4.71	(0.67)
Item 17	Camisa	2.39	(0.30)	1.96	(0.24)
Item 18	Cansado	3.03	(0.36)	0.77	(0.22)
Item 19	Carne	2.26	(0.29)	2.21	(0.25)
Item 20	Carro	2.79	(0.42)	4.13	(0.50)
Item 21	Cómo	2.94	(0.34)	-0.03	(0.21)
Item 22	Comprar	3.06	(0.36)	0.27	(0.21)
Item 23	Culebra/serpiente	2.31	(0.30)	-1.35	(0.21)
Item 24	Dónde	3.02	(0.36)	0.77	(0.22)
Item 25	Dormir(se)	3.10	(0.43)	3.34	(0.41)
Item 26	En la mañana	2.58	(0.32)	-1.37	(0.23)
Item 27	Escoba	3.14	(0.38)	1.55	(0.25)
Item 28	Estar	3.36	(0.40)	-0.42	(0.23)
Item 29	Falda	2.98	(0.36)	-0.98	(0.23)
Item 30	Fiesta	3.16	(0.37)	0.50	(0.22)
Item 31	Flor	3.28	(0.40)	1.54	(0.26)
Item 32	Fósforos	1.76	(0.25)	-1.74	(0.21)
Item 33	Ganar	2.87	(0.34)	-0.61	(0.21)
Item 34	Gato	3.28	(0.47)	3.62	(0.45)
Item 35	Grande	2.54	(0.31)	1.38	(0.22)
Item 36	guaguáu	1.42	(0.28)	3.94	(0.43)
Item 37	Haber (hay)	1.43	(0.19)	0.74	(0.15)
Item 38	Hacer	2.79	(0.33)	0.31	(0.20)
Item 39	Hoy	2.50	(0.29)	0.64	(0.19)
Item 40	Huevo	2.23	(0.29)	2.43	(0.27)
Item 41	Iglesia	3.24	(0.39)	-1.08	(0.25)
Item 42	Jabón	3.44	(0.47)	3.02	(0.39)
Item 43	Jugar	4.36	(0.68)	4.51	(0.64)
Item 44	La	1.19	(0.18)	1.88	(0.19)
Item 45	Leche	1.97	(0.27)	2.65	(0.28)
Item 46	Libro	2.77	(0.34)	1.61	(0.24)
Item 47	Llover	2.91	(0.35)	0.82	(0.22)
Item 48	Luz	2.28	(0.32)	3.16	(0.35)
Item 49	Madrina	1.78	(0.24)	-1.15	(0.18)
Item 50	Malo	3.11	(0.37)	0.75	(0.22)
Item 51	Mamá	1.14	(0.50)	5.54	(0.90)
Item 52	Manguera	2.82	(0.39)	-2.51	(0.32)
Item 53	Mano	3.26	(0.52)	5.00	(0.67)
Item 54	Mirar	2.28	(0.28)	1.59	(0.22)
Item 55	Mío	1.98	(0.32)	4.13	(0.48)
Item 56	Más	1.47	(0.23)	2.82	(0.28)
Item 57	Muu	1.04	(0.21)	2.80	(0.26)
Item 58	Niño	2.00	(0.32)	3.97	(0.45)
Item 59	No hay	2.43	(0.29)	1.35	(0.21)
Item 60	Nuestro	2.61	(0.35)	-1.95	(0.26)
Item 61	Nuevo	3.30	(0.39)	-0.57	(0.23)
Item 62	Oír	2.35	(0.28)	1.02	(0.20)
Item 63	Olla	2.50	(0.30)	0.66	(0.19)
Item 64	Pantalón	2.69	(0.34)	2.17	(0.27)
Item 65	Papas	2.08	(0.31)	3.44	(0.38)
Item 66	Pato	2.36	(0.29)	1.53	(0.22)
Item 67	Payaso	2.26	(0.27)	-0.36	(0.18)
Item 68	Pelota	2.49	(0.39)	4.36	(0.53)

Panel H: MacArthur-Bates - Older Kids

Item (j)	Word	β_j		α_j	
Item 1	adelante	2.08	(0.32)	2.04	(0.26)
Item 2	ambulancia	1.31	(0.20)	0.19	(0.15)
Item 3	aquel	1.02	(0.17)	-0.16	(0.14)
Item 4	arreglar	1.62	(0.26)	1.76	(0.22)
Item 5	atrás	1.85	(0.33)	3.09	(0.36)
Item 6	ayer	1.31	(0.21)	0.83	(0.16)
Item 7	barba	1.23	(0.19)	-0.59	(0.15)
Item 8	biblioteca (pública)	1.87	(0.27)	-1.35	(0.21)
Item 9	bolsa	1.53	(0.31)	3.21	(0.36)
Item 10	caber	1.55	(0.23)	0.35	(0.17)
Item 11	cada	1.60	(0.23)	-0.17	(0.17)
Item 12	candado	1.24	(0.20)	0.23	(0.15)
Item 13	cesta o canasta	2.17	(0.31)	0.99	(0.21)
Item 14	clínica o hospital	1.69	(0.24)	0.37	(0.17)
Item 15	computadora	1.58	(0.27)	2.26	(0.26)
Item 16	contra	1.53	(0.23)	-1.15	(0.18)
Item 17	cuadrado	1.55	(0.24)	1.21	(0.19)
Item 18	cueva	2.19	(0.30)	-1.11	(0.22)
Item 19	cuál	1.44	(0.25)	1.84	(0.22)
Item 20	dañado	1.59	(0.32)	3.25	(0.37)
Item 21	descansar	2.43	(0.35)	1.72	(0.26)
Item 22	después	2.03	(0.30)	1.45	(0.22)
Item 23	dinosaurio	1.29	(0.21)	0.72	(0.16)
Item 24	echar	1.72	(0.25)	0.34	(0.17)
Item 25	empujar	1.91	(0.32)	2.61	(0.31)
Item 26	enfermo	1.79	(0.32)	2.86	(0.33)
Item 27	escalera	1.92	(0.36)	3.54	(0.42)
Item 28	estantería o armario	1.66	(0.24)	-0.32	(0.17)
Item 29	fábrica	2.32	(0.32)	0.46	(0.21)
Item 30	faltar	2.82	(0.39)	-1.70	(0.28)
Item 31	figura	2.66	(0.38)	1.47	(0.26)
Item 32	flecha	1.87	(0.26)	-0.09	(0.18)
Item 33	garganta	1.84	(0.27)	1.01	(0.20)
Item 34	grupo	2.95	(0.40)	-0.07	(0.24)
Item 35	hasta	2.39	(0.33)	0.22	(0.21)
Item 36	herramienta	1.78	(0.25)	-0.11	(0.17)
Item 37	horno	1.72	(0.24)	-0.11	(0.17)
Item 38	idea	2.23	(0.30)	-0.22	(0.20)
Item 39	igual	2.58	(0.36)	1.26	(0.25)
Item 40	insecto	1.41	(0.21)	-0.29	(0.16)
Item 41	jalar	1.41	(0.22)	1.14	(0.18)
Item 42	juntar	1.84	(0.27)	0.95	(0.19)
Item 43	lado	2.36	(0.36)	2.16	(0.29)
Item 44	lastimar	2.04	(0.30)	1.63	(0.23)
Item 45	letra	2.35	(0.35)	1.88	(0.27)
Item 46	línea	2.42	(0.34)	1.01	(0.23)
Item 47	lugar	2.97	(0.43)	1.83	(0.30)
Item 48	manejar	1.46	(0.24)	1.37	(0.19)
Item 49	mecánico	1.68	(0.25)	-1.48	(0.21)
Item 50	medir	2.97	(0.40)	-0.55	(0.24)
Item 51	meter	1.76	(0.28)	1.99	(0.24)
Item 52	mis	1.19	(0.21)	1.59	(0.19)
Item 53	mismo	1.78	(0.26)	0.64	(0.18)
Item 54	montaña	2.09	(0.30)	0.97	(0.21)
Item 55	mover	2.34	(0.37)	2.71	(0.34)
Item 56	mueble	1.44	(0.25)	1.95	(0.23)
Item 57	muy	1.73	(0.27)	1.60	(0.22)
Item 58	necesitar	2.36	(0.33)	0.60	(0.21)
Item 59	nido	2.04	(0.28)	-0.50	(0.19)
Item 60	nosotros	2.66	(0.39)	2.12	(0.30)
Item 61	oficina	2.05	(0.28)	-0.88	(0.20)
Item 62	oscuro	1.62	(0.28)	2.25	(0.26)
Item 63	parecer	3.27	(0.45)	-0.66	(0.26)
Item 64	peligroso	2.95	(0.40)	0.65	(0.24)
Item 65	(pelo) corto	1.03	(0.21)	1.91	(0.20)
Item 66	pequeño	2.73	(0.46)	3.66	(0.48)
Item 67	pera	1.90	(0.29)	1.86	(0.24)
Item 68	perder	2.11	(0.30)	0.98	(0.21)

Item 69	Periódico	2.55	(0.34)	-2.14	(0.27)	Item 69	perfecto	2.52	(0.35)	-1.24	(0.24)
Item 70	Plátano/banano	2.29	(0.28)	1.37	(0.21)	Item 70	perseguir	2.48	(0.34)	0.32	(0.21)
Item 71	Pollo	3.15	(0.45)	3.54	(0.44)	Item 71	persona	2.62	(0.39)	2.27	(0.31)
Item 72	Por favor	2.41	(0.29)	1.28	(0.21)	Item 72	pesado	1.71	(0.28)	2.15	(0.25)
Item 73	Prender	2.11	(0.25)	0.56	(0.17)	Item 73	pintor	1.73	(0.24)	-0.41	(0.17)
Item 74	Puerta	3.42	(0.48)	3.22	(0.41)	Item 74	plástico	2.57	(0.35)	0.69	(0.22)
Item 75	Quién	2.83	(0.33)	0.59	(0.21)	Item 75	por	1.58	(0.25)	1.29	(0.19)
Item 76	Quiquiriquí	1.46	(0.19)	-0.50	(0.15)	Item 76	pulsera	1.68	(0.25)	1.12	(0.19)
Item 77	Rana	1.89	(0.23)	-0.18	(0.16)	Item 77	puntilla	1.64	(0.23)	-0.25	(0.17)
Item 78	Roto	2.35	(0.28)	0.66	(0.19)	Item 78	quedar	1.61	(0.24)	0.91	(0.18)
Item 79	Sí	2.30	(0.45)	5.36	(0.78)	Item 79	raqueta	1.36	(0.21)	-0.31	(0.16)
Item 80	Saber	2.85	(0.34)	-0.67	(0.21)	Item 80	raro	2.74	(0.37)	0.02	(0.22)
Item 81	Saltar	2.21	(0.27)	1.39	(0.20)	Item 81	regresar	2.48	(0.34)	0.18	(0.21)
Item 82	Sentar(se)	3.33	(0.45)	3.16	(0.40)	Item 82	río	2.45	(0.37)	2.18	(0.29)
Item 83	Sol	2.58	(0.32)	2.03	(0.26)	Item 83	saber	2.82	(0.40)	1.56	(0.27)
Item 84	Ésta	2.07	(0.25)	0.76	(0.18)	Item 84	salvar	2.70	(0.36)	-0.43	(0.22)
Item 85	Sucio	3.10	(0.42)	3.08	(0.38)	Item 85	sembrar	2.49	(0.34)	-0.37	(0.21)
Item 86	Suya	2.10	(0.25)	0.71	(0.18)	Item 86	semilla	2.25	(0.31)	-0.54	(0.20)
Item 87	Tambor	2.37	(0.29)	-0.82	(0.19)	Item 87	sobre (la silla)	1.90	(0.29)	1.55	(0.22)
Item 88	Televisión	2.56	(0.32)	2.05	(0.26)	Item 88	sus	1.74	(0.26)	1.34	(0.20)
Item 89	Tetero	0.92	(0.20)	2.76	(0.25)	Item 89	suyos	1.30	(0.23)	1.69	(0.20)
Item 90	Tigre	1.89	(0.23)	-0.08	(0.16)	Item 90	también	1.89	(0.31)	2.29	(0.27)
Item 91	Timbre	2.35	(0.29)	-1.07	(0.20)	Item 91	ti	0.98	(0.19)	1.19	(0.16)
Item 92	Tomate	2.23	(0.27)	0.26	(0.18)	Item 92	tigre	1.67	(0.27)	1.95	(0.24)
Item 93	Tutu	0.78	(0.14)	0.17	(0.12)	Item 93	torre	2.42	(0.34)	0.97	(0.22)
Item 94	Vaca	2.77	(0.35)	2.33	(0.29)	Item 94	tractor	1.63	(0.24)	-1.11	(0.19)
Item 95	Vasos	2.41	(0.33)	2.82	(0.32)	Item 95	tranquilo	2.53	(0.35)	0.77	(0.22)
Item 96	Vámonos	2.23	(0.33)	3.64	(0.41)	Item 96	vainilla	1.88	(0.26)	-0.90	(0.19)
Item 97	Zapato	2.53	(0.37)	3.97	(0.47)	Item 97	vender	2.59	(0.35)	0.50	(0.22)
						Item 98	verdura	1.57	(0.24)	1.21	(0.19)

Notes: Table presents estimated parameters for IRT measurement model of children's development measured at baseline alongside estimated standard errors in parentheses. Standard errors are calculated analytically since we do not bootstrap the baseline measurement systems. 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.5: Measurement Model Parameters: Teacher Learning Activities

Item (j)	β_j		α_{j1}		α_{j2}		α_{j3}		α_{j4}		α_{j5}	
Number of times did ..... last week												
Read stories	1.31	[.95, 1.69]	5.02	[4.64, 5.66]	2.79	[2.37, 3.41]	1.71	[1.35, 2.2]	0.74	[.41, 1.15]	0.48	[.15, .85]
Tell stories	1.41	[1.18, 1.63]	3.59	[3.13, 4.22]	2.04	[1.72, 2.46]	1.05	[.73, 1.42]	0.19	[-.12, .51]	-0.04	[-.35, .29]
Conversation	2.54	[1.75, 4.79]	8.33	[6.65, 14.43]	7.43	[6.53, 13.13]	6.31	[5.59, 10.33]	5.98	[5.33, 9.13]	5.01	[4.41, 7.62]
Sing	1.73	[.79, 3.3]	7.67	[6.45, 12.72]	6.38	[5.45, 7.02]	5.74	[4.63, 9.03]	4.48	[3.54, 6.91]	3.80	[3.07, 5.7]
Dance	0.87	[.62, 1.21]	3.54	[3.14, 4.24]	2.49	[2.13, 3.03]	1.29	[1.01, 1.65]	0.68	[.39, 1.05]	0.52	[.25, .88]
Watch a video or educational programme on TV.	0.46	[.24, .68]	0.48	[.16, .82]	-0.83	[-1.15, -.53]	-1.73	[-2.11, -1.41]	-2.64	[-3.05, -2.25]	-2.76	[-3.23, -2.41]
isit other places in the community	0.36	[.01, .68]	-1.53	[-1.89, -1.24]	-3.52	[-4.43, -3.02]	-4.52	[-5.01, -3.65]		[., .]	-5.63	[-5.73, -4.55]
Free play within the nursery premises	0.86	[.56, 1.2]	4.54	[3.81, 5.82]	2.81	[2.43, 3.33]	1.64	[1.31, 2.04]	1.16	[.86, 1.51]	1.01	[.72, 1.36]
Free play in the recreation area	0.50	[.25, .75]	1.89	[1.53, 2.26]	0.79	[.52, 1.08]	0.18	[-.12, .46]	-0.34	[-.63, -.06]	-0.42	[-.72, -.14]
Physical activities such as running, jumping	0.84	[.59, 1.09]	4.52	[3.99, 5.33]	2.31	[1.99, 2.68]	1.03	[.79, 1.31]	0.50	[.23, .79]	0.36	[.09, .65]
Group learning activities	1.00	[.61, 1.43]	3.90	[3.45, 4.59]	2.88	[2.51, 3.38]	2.16	[1.82, 2.65]	1.40	[1.1, 1.79]	1.06	[.75, 1.43]
Individual learning activities	1.36	[1.05, 1.69]	3.41	[2.98, 4]	2.54	[2.17, 3.02]	1.83	[1.5, 2.25]	1.20	[.9, 1.55]	0.79	[.49, 1.14]
Teach colours	1.36	[1.12, 1.68]	1.82	[1.55, 2.18]	1.34	[1.11, 1.67]	0.68	[.46, 1]	0.30	[.06, .62]	0.22	[-.01, .53]
Teach numbers	0.96	[.63, 1.32]	0.58	[.35, .84]	0.27	[.05, .51]	-0.13	[-.36, .09]	-0.47	[-.7, -.24]	-0.56	[-.79, -.35]
Teach letter of the alphabet	0.76	[.39, 1.12]	-0.27	[-.53, -.03]	-0.60	[-.87, -.35]	-1.04	[-1.32, -.82]	-1.20	[-1.48, -.98]	-1.35	[-1.62, -1.13]
Teach forms and shapes	1.74	[1.37, 2.19]	1.94	[1.6, 2.42]	1.09	[.81, 1.46]	0.38	[.11, .72]	-0.14	[-.42, .21]	-0.42	[-.72, -.08]
Socialising	1.99	[1.59, 2.7]	7.10	[5.74, 9.63]	4.74	[4.06, 6.32]	4.31	[3.72, 5.62]	3.89	[3.4, 4.92]	3.39	[2.94, 4.26]
Problem solving	1.32	[1.02, 1.67]	3.05	[2.69, 3.61]	2.70	[2.33, 3.21]	2.13	[1.82, 2.56]	1.78	[1.45, 2.21]	1.59	[1.27, 2.01]
Writing	0.87	[.52, 1.25]	0.26	[-.01, .51]	-0.13	[-.38, .1]	-0.45	[-.74, -.19]	-1.06	[-1.37, -.79]	-1.18	[-1.49, -.93]
Teach parts of the body	1.52	[1.21, 1.95]	2.42	[2.07, 2.93]	1.62	[1.31, 2.08]	0.97	[.66, 1.39]	0.24	[-.04, .62]	0.02	[-.24, .37]
Teach about personal hygiene and body care	1.51	[1.23, 1.93]	4.98	[4.32, 6.23]	3.47	[3.02, 4.25]	3.18	[2.77, 3.88]	2.75	[2.39, 3.37]	2.49	[2.16, 3.05]
Artistic expression	1.25	[.99, 1.6]	3.29	[2.9, 3.89]	1.44	[1.2, 1.77]	0.61	[.37, .93]	-0.20	[-.45, .13]	-0.38	[-.65, -.07]
Body language	1.57	[1.31, 1.9]	4.79	[4.24, 5.71]	2.48	[2.14, 2.96]	1.32	[1, 1.72]	0.88	[.59, 1.24]	0.68	[.36, 1.06]
Concentration	2.01	[1.7, 2.4]	2.87	[2.46, 3.4]	2.20	[1.82, 2.71]	1.41	[1.05, 1.88]	0.81	[.46, 1.25]	0.51	[.15, .94]
Gross motor coordination	1.26	[.97, 1.62]	4.44	[3.96, 5.24]	2.72	[2.4, 3.19]	1.63	[1.31, 2.05]	0.81	[.52, 1.18]	0.61	[.31, 1]
Fine motor coordination	1.58	[1.28, 1.91]	3.87	[3.37, 4.56]	2.72	[2.41, 3.13]	1.80	[1.48, 2.2]	0.93	[.61, 1.29]	0.60	[.29, .92]
Gender identification	1.86	[1.55, 2.3]	2.37	[2.01, 2.92]	1.74	[1.38, 2.26]	1.15	[.81, 1.64]	0.89	[.55, 1.36]	0.71	[.38, 1.17]
Responsibility	1.78	[1.38, 2.22]	3.48	[3.07, 4.04]	2.70	[2.32, 3.21]	2.26	[1.92, 2.72]	1.99	[1.64, 2.45]	1.80	[1.48, 2.25]
Speech/story telling	1.60	[1.22, 2.06]	3.04	[2.61, 3.56]	2.31	[1.98, 2.76]	1.59	[1.25, 2.01]	1.05	[.69, 1.5]	0.91	[.59, 1.32]
Explore the community	1.01	[.73, 1.26]	-0.81	[-1.09, -.53]	-1.70	[-2.03, -1.42]	-2.63	[-2.99, -2.34]	-2.99	[-3.39, -2.65]	-3.14	[-3.58, -2.78]
Explore regional culture	1.38	[1.12, 1.65]	-0.35	[-.59, -.07]	-1.65	[-1.98, -1.34]	-2.48	[-2.89, -2.11]	-2.84	[-3.25, -2.46]	-3.00	[-3.47, -2.61]
Describe ..... as being a main duty												
Plan pedagogic component of all daily activities.	0.07	[-.27, .48]	1.18	[.9, 1.64]								
Implement educational and social activities with children.	0.21	[-.01, .48]	1.17	[.87, 1.56]								
Provide quality attention.	0.34	[.13, .6]	1.04	[.81, 1.31]								
Teach norms, limits and group agreements	0.20	[-.06, .53]	0.47	[.21, .78]								
Strength verbal and body language.	0.16	[-.07, .49]	-0.06	[-.39, .24]								

Notes: Table presents estimated parameters for IRT measurement model of teachers' learning activities at endline alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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.6: Measurement Model Parameters: Teacher Care Activities

Item (j)	β_j		α_{j1}		α_{j2}		α_{j3}		α_{j4}		α_{j5}	
Number of times did ..... last week												
Hygiene routines and care, e.g. changing nappies, brush teeth, wash hands	0.42	[.28, .62]	-0.07	[-.68, 1.44]			0.36	[-.51, 1.53]	2.22	[.48, 3.02]	6.21	[4.83, 7.11]
Supply medicines / remedies	0.42	[.28, .62]	-3.12	[-4.17, -2.51]	-3.39	[-4.12, -2.93]	-4.10	[-4.93, -3.63]			-3.69	[-4.67, -3.06]
Pamper children	0.42	[.28, .63]	1.46	[.29, 2.28]	2.05	[0, 2.51]	1.76	[-.12, 2.47]	3.79	[1.14, 4.2]	7.58	[5.26, 8.07]
Watch TV with children	0.42	[.28, .62]	-2.51	[-3, -2.12]	-3.10	[-3.88, -2.56]	-3.49	[-4.59, -2.78]			-3.57	[-4.93, -2.77]
Describe ..... as being a main duty												
Teach and support children during practices on personal hygiene.	0.42	[.28, .63]	1.39	[1.08, 1.75]								
Support during meals.	0.42	[.28, .63]	1.57	[1.23, 1.97]								

Notes: Table presents estimated parameters for IRT measurement model of teachers' caring activities at endline alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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.7: Measurement Model Parameters: TA Learning Activities

Item (j)	β_j		α_{j1}		α_{j2}		α_{j3}		α_{j4}		α_{j5}	
Number of times did ..... last week												
Read stories	0.70	[.32, 1.07]	2.34	[1.98, 2.75]	1.54	[1.13, 1.98]	0.08	[-.26, .35]	-0.63	[-.98, -.32]	-0.86	[-1.23, -.56]
Tell stories	1.18	[.89, 1.52]	1.40	[1.01, 1.79]	0.46	[.13, .78]	-0.46	[-.82, -.12]	-1.02	[-1.39, -.7]	-1.28	[-1.68, -.96]
Conversation	2.07	[1.4, 3.15]		[., .]	5.65	[4.67, 7.52]	4.97	[4.05, 7.03]	4.71	[3.78, 6.38]	2.86	[2.17, 4.07]
Sing	1.67	[.88, 2.73]		[., .]	6.24	[5.46, 7.97]		[., .]	4.93	[4.03, 6.77]	2.71	[2.08, 3.72]
Dance	1.07	[.56, 1.56]	3.00	[2.46, 3.7]	2.51	[1.95, 3.12]	1.77	[1.24, 2.3]	0.83	[.5, 1.14]	0.38	[.07, .7]
Watch a video or educational programme on TV.	0.41	[.14, .68]	0.35	[0, .7]	-1.24	[-1.66, -.87]	-2.05	[-2.61, -1.65]	-2.89	[-3.64, -2.31]	-3.05	[-4.11, -2.48]
isit other places in the community	0.55	[.19, .93]	2.81	[2.38, 3.47]	1.58	[1.3, 1.97]	1.06	[.76, 1.44]	0.59	[.27, .91]	0.29	[-.03, .63]
Free play within the nursery premises	0.43	[.09, .77]	1.63	[1.3, 2.01]	0.87	[.54, 1.24]	0.12	[-.25, .43]	-0.38	[-.77, -.08]	-0.70	[-1.12, -.4]
Free play in the recreation area	0.79	[.51, 1.15]	2.95	[2.58, 3.57]	1.42	[1.11, 1.78]	0.77	[.43, 1.1]	0.14	[-.22, .45]	-0.23	[-.62, .1]
Physical activities such as running, jumping	1.75	[1.33, 2.24]	3.01	[2.52, 3.7]	1.88	[1.36, 2.38]	1.05	[.6, 1.53]	0.58	[.12, .97]	0.28	[-.14, .69]
Group learning activities	1.34	[1, 1.74]	1.35	[.88, 1.78]	0.93	[.45, 1.35]	0.37	[-.13, .79]	0.07	[-.43, .46]	-0.14	[-.62, .23]
Teach colours	1.70	[1.17, 2.57]	0.38	[-.09, .78]	-0.23	[-.77, .15]	-0.60	[-1.21, -.2]	-0.98	[-1.58, -.6]	-1.24	[-1.85, -.87]
Teach numbers	1.18	[.77, 1.82]	-0.06	[-.47, .26]	-0.64	[-1.05, -.35]	-0.82	[-1.23, -.58]	-1.11	[-1.54, -.8]	-1.27	[-1.75, -.98]
Teach letter of the alphabet	1.14	[.64, 1.89]	-1.30	[-1.75, -1]	-1.74	[-2.24, -1.47]	-2.25	[-2.82, -1.95]	-2.52	[-3.07, -2.19]	-2.63	[-3.31, -2.33]
Teach forms and shapes	1.42	[1.02, 2.04]	1.02	[.65, 1.38]	-0.02	[-.4, .29]	-0.61	[-.99, -.35]	-1.15	[-1.6, -.85]	-1.42	[-1.88, -1.11]
Socialising	1.65	[1.12, 2.34]	4.38	[3.6, 5.55]	2.95	[2.26, 3.83]	2.39	[1.8, 3.07]	2.15	[1.64, 2.75]	1.53	[1.06, 2.02]
Problem solving	1.56	[1.09, 2.17]	1.83	[1.4, 2.29]	1.38	[.95, 1.8]	1.20	[.81, 1.58]	1.04	[.68, 1.43]	0.74	[.4, 1.08]
Writing	0.95	[.66, 1.35]	-0.46	[-.75, -.2]	-0.95	[-1.31, -.68]	-1.43	[-1.83, -1.11]	-1.90	[-2.34, -1.58]	-2.18	[-2.65, -1.85]
Teach parts of the body	1.01	[.68, 1.41]	0.61	[.21, .91]	-0.11	[-.5, .21]	-0.66	[-1.05, -.37]	-0.96	[-1.31, -.63]	-1.09	[-1.51, -.79]
Teach about personal hygiene and body care	1.83	[1.2, 2.77]	3.30	[2.6, 4.42]	2.80	[2.13, 3.77]	2.51	[1.83, 3.42]	2.19	[1.51, 3.13]	1.56	[.97, 2.3]
Artistic expression	0.72	[.47, 1.07]	1.92	[1.56, 2.39]	0.35	[.1, .69]	-0.14	[-.42, .22]	-0.80	[-1.11, -.51]	-1.01	[-1.34, -.7]
Body language	0.96	[.66, 1.35]	2.27	[1.8, 2.88]	0.75	[.48, 1.12]	0.01	[-.31, .32]	-0.18	[-.51, .07]	-0.38	[-.74, -.08]
Concentration	0.99	[.68, 1.47]	1.16	[.79, 1.52]	0.15	[-.22, .53]	-0.32	[-.67, 0]	-0.60	[-.99, -.26]	-0.80	[-1.22, -.43]
Gross motor coordination	1.09	[.84, 1.57]	2.51	[2.09, 3.12]	1.06	[.71, 1.41]	0.35	[-.02, .69]	0.02	[-.4, .34]	-0.19	[-.63, .13]
Fine motor coordination	1.31	[.97, 1.99]	2.63	[2.23, 3.35]	1.27	[.9, 1.73]	0.34	[0, .7]	-0.31	[-.69, .03]	-0.62	[-1.04, -.28]
Gender identification	1.25	[.95, 1.71]	0.70	[.29, 1.15]	0.00	[-.32, .3]	-0.26	[-.56, .05]	-0.43	[-.74, -.15]	-0.65	[-.96, -.37]
Responsibility	1.27	[.93, 1.71]	1.86	[1.43, 2.36]	1.35	[.99, 1.76]	1.25	[.88, 1.68]	1.06	[.68, 1.47]	0.59	[.22, 1]
Speech/story telling	1.56	[1.2, 2.01]	1.40	[.93, 1.82]	0.78	[.36, 1.23]	0.14	[-.28, .49]	-0.35	[-.74, .02]	-0.53	[-.96, -.13]
Explore the community	0.46	[.1, .85]	-1.22	[-1.69, -.91]	-2.29	[-2.82, -2.03]	-3.05	[-3.63, -2.65]	-3.24	[-3.97, -2.79]	-3.48	[-4.62, -3.03]
Explore regional culture	0.45	[.17, .75]	-1.29	[-1.76, -.97]	-1.99	[-2.64, -1.61]	-2.63	[-3.37, -2.18]	-3.06	[-3.54, -2.5]	-3.49	[-4.55, -3.01]
Describe ..... as being a main duty												
Prepare in assessment reports about children	-0.22	[-.64, .14]	-1.85	[-2.27, -1.51]								
Prepare in descriptive reports about classroom activities	-0.07	[-.33, .24]	-2.05	[-2.6, -1.69]								
Support the development of goals at the centre	-0.08	[-.62, .47]	-2.32	[-2.93, -1.96]								
Contribute to design and implementation of teaching strategies	0.47	[.17, .82]	-0.58	[-1.02, -.3]								
Perform work with families of the children	-0.04	[-.35, .26]	-4.10	[-4.29, -3.33]								
Organise teaching materials	0.16	[-.22, .59]	-0.70	[-.96, -.41]								

Notes: Table presents estimated parameters for IRT measurement model of TAs' learning activities at endline alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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.8: Measurement Model Parameters: TA Care Activities

Item (j)		β_j		α_j
1 if Number of times did ..... last week is more than mean for TAs, 0 otherwise				
Hygiene routines and care such as changing nappies, brush teeth, wash hands	0.85	[.35, 1.35]	1.80	[1.36, 2.43]
Supply medicines / remedies	0.85	[.35, 1.35]	-3.48	[-4.45, -2.96]
Pamper children	0.85	[.35, 1.35]	2.00	[1.52, 2.66]
Watch TV with children	0.85	[.35, 1.35]	-2.50	[-3.16, -2.05]
Describe ..... as being a main duty				
Perform personal hygiene activities with the children such as changing nappies, wash hands, feeding during meals, care during rest time, etc	0.85	[.35, 1.35]	1.86	[1.35, 2.63]
Clean the classroom	0.85	[.35, 1.35]	-4.45	[-5.43, -3.86]

Notes: Table presents estimated parameters for IRT measurement model of TAs' caring activities at endline alongside 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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.9: Measurement Model Parameters: Baseline Teacher Learning Activities

Item (j)	β_j		α_{j1}		α_{j2}		α_{j3}		α_{j4}		α_{j5}	
Number of times did ..... last week												
Read stories	0.80	(0.16)	4.40	(0.52)	2.70	(0.25)	1.70	(0.19)	0.76	(0.15)	0.55	(0.15)
Tell stories	1.25	(0.19)	3.24	(0.31)	2.30	(0.23)	1.17	(0.18)	0.44	(0.17)	0.28	(0.17)
Conversation	2.94	(0.81)			8.72	(1.79)	6.92	(1.32)	6.62	(1.27)	5.69	(1.14)
Sing	1.55	(0.44)	6.62	(1.16)	5.90	(0.90)	5.44	(0.78)	4.85	(0.66)	3.66	(0.51)
Dance	0.88	(0.17)	4.03	(0.43)	2.78	(0.26)	1.43	(0.18)	0.75	(0.16)	0.55	(0.15)
Watch a video or educational programme on TV.	-0.16	(0.16)	-1.23	(0.15)	-1.83	(0.18)	-2.77	(0.27)	-3.56	(0.38)	-3.72	(0.41)
isit other places in the community	0.45	(0.18)	-1.59	(0.18)	-3.09	(0.31)	-3.51	(0.37)			-4.52	(0.59)
Free play within the nursery premises	0.98	(0.20)	3.96	(0.41)	2.81	(0.27)	1.90	(0.21)	1.51	(0.19)	1.23	(0.18)
Free play in the recreation area	0.89	(0.16)	2.11	(0.21)	1.32	(0.17)	0.49	(0.15)	0.02	(0.15)	-0.24	(0.15)
Physical activities such as running, jumping	1.22	(0.21)	4.11	(0.43)	2.88	(0.28)	1.60	(0.20)	1.12	(0.18)	0.71	(0.18)
Group learning activities	1.59	(0.29)	4.89	(0.56)	4.04	(0.43)	3.46	(0.37)	2.54	(0.30)	1.94	(0.27)
Teach colours	1.02	(0.19)	2.95	(0.28)	2.61	(0.26)	2.05	(0.22)	1.40	(0.19)	1.04	(0.18)
Teach numbers	1.13	(0.18)	2.01	(0.21)	1.22	(0.18)	0.71	(0.17)	0.32	(0.16)	0.09	(0.16)
Teach letter of the alphabet	0.85	(0.16)	0.00	(0.15)	-0.18	(0.15)	-0.51	(0.15)	-0.79	(0.15)	-0.92	(0.16)
Teach forms and shapes	0.73	(0.17)	-0.86	(0.15)	-1.12	(0.16)	-1.51	(0.18)	-1.84	(0.20)	-1.95	(0.20)
Socialising	1.48	(0.20)	2.34	(0.24)	1.67	(0.21)	0.79	(0.18)	0.16	(0.18)	-0.18	(0.18)
Problem solving	1.82	(0.34)	5.52	(0.67)	4.59	(0.54)	3.95	(0.47)	3.55	(0.43)	2.79	(0.37)
Writing	0.94	(0.16)	-0.13	(0.15)	-0.34	(0.15)	-0.86	(0.16)	-1.39	(0.18)	-1.59	(0.19)
Teach parts of the body	1.42	(0.22)	2.93	(0.29)	2.14	(0.24)	1.48	(0.21)	1.04	(0.19)	0.80	(0.19)
Teach about personal hygiene and body care	2.78	(0.53)	6.45	(0.90)	6.05	(0.85)	5.14	(0.73)	4.59	(0.66)	3.82	(0.59)
Artistic expression	1.33	(0.19)	3.45	(0.33)	2.02	(0.22)	0.94	(0.18)	0.26	(0.17)	-0.05	(0.17)
Body language	1.32	(0.21)	4.39	(0.46)	2.23	(0.24)	1.40	(0.20)	1.07	(0.19)	0.74	(0.18)
Concentration	1.48	(0.22)	2.29	(0.25)	1.84	(0.22)	1.41	(0.21)	0.80	(0.19)	0.55	(0.18)
Gross motor coordination	1.38	(0.24)	4.90	(0.56)	3.78	(0.38)	2.65	(0.28)	1.65	(0.22)	1.38	(0.21)
Fine motor coordination	1.40	(0.21)	4.70	(0.50)	3.38	(0.32)	2.26	(0.24)	1.23	(0.20)	0.72	(0.19)
Gender identification	1.64	(0.25)	2.73	(0.29)	2.09	(0.25)	1.59	(0.23)	1.21	(0.22)	0.98	(0.21)
Responsibility	1.97	(0.32)	3.19	(0.37)	2.90	(0.35)	2.57	(0.33)	2.29	(0.31)	1.97	(0.29)
Speech/story telling	1.42	(0.24)	3.74	(0.38)	2.93	(0.30)	2.25	(0.26)	1.91	(0.24)	1.50	(0.22)
Explore the community	0.96	(0.17)	-0.39	(0.15)	-1.23	(0.17)	-1.77	(0.20)	-2.12	(0.22)	-2.48	(0.24)
Explore regional culture	0.96	(0.17)	-0.54	(0.16)	-1.36	(0.18)	-1.86	(0.20)	-2.19	(0.22)	-2.23	(0.22)

Notes: Table presents estimated parameters for IRT measurement model of teachers' learning activities at endline alongside estimated standard errors in parentheses. Standard errors are calculated analytically since we do not bootstrap the baseline measurement systems. 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.10: Measurement Model Parameters: Baseline Teacher Care Activities

Item (j)	β_j		α_{j1}		α_{j2}		α_{j3}		α_{j4}		α_{j5}	
Number of times did ..... last week												
Hygiene routines and care such as changing nappies, brush teeth, wash hands	0.48	(0.24)			3.66	(0.40)	0.36	(0.78)	2.22	(0.97)	6.21	(0.85)
Supply medicines / remedies	0.48	(0.24)	-1.87	(0.20)	-2.54	(0.37)	-4.10	(0.49)			-3.69	(0.65)
Pamper children	0.48	(0.24)			3.29	(0.34)	1.76	(0.99)	3.79	(1.14)	7.58	(1.05)
Watch TV with children	0.48	(0.24)	-1.39	(0.17)	-3.80	(0.52)	-3.49	(0.71)			-3.57	(0.86)

Notes: Table presents estimated parameters for IRT measurement model of teachers' caring activities at endline alongside estimated standard errors in parentheses. Standard errors are calculated analytically since we do not bootstrap the baseline measurement systems. 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.11: Measurement Model Parameters: ECERS-R Diect 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 90% confidence intervals in brackets. Confidence intervals are constructed using our block bootstrap described in Section 3.3, resampling triplets with replacements (1000 iterations). 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 Modelling Appendix

In this appendix we provide additional explanation for the model results presented in the main text. We first discuss the more-general version of our model that we work with in Sections 6.1 and 6.2. We then discuss our more-specialized version that we work with in Section 6.3.

D.1 Appendix to Generalized Model

Here we show how to derive equations (6.5) and (6.5) from the main text which describe how teachers' choices of care and learning activities change with the addition of TA time. As outlined in 6.1 we assume that child development H is produced by combining teachers' time spent doing learning activities (L_t), teachers' time spent doing personal care activities (C_t), and teaching assistants' time (A). These inputs are combined through the following production function:

$$H = zf(L_t, C_t, A)$$

Let the derivatives of the production function with respect to each argument be denoted f_1, f_2 and f_3, and the second and cross derivatives by, for instance, f_{11} and f_{12}. We assume that this production function is continuous, is increasing in all arguments and is concave. We assume that any q-substitutability between teachers' learning and care activities quantitatively smaller than the rate at which the marginal product of these inputs diminishes, i.e. $f_{12} > \max(f_{11}, f_{22})$.

As outlined in the main text, teachers' utility is given by $u(H, K)$ where $K = 1 - L_t - C_t$. The utility function increasing increasing in both arguments, concave and is separable in child development and leisure (i.e. $u_{HK} = 0$).

Teachers choose C_t and L_t taking A as given, subject to the production function. The teachers' problem is thus:

$$\max_{L_t, C_t} u(zf(L_t, C_t, A), 1 - L_t - C_t) \tag{D.1}$$

which leads to the following first-order conditions:

$$0 = zu_H f_1 - u_K \tag{D.2}$$

$$0 = zu_H f_2 - u_K \tag{D.3}$$

We use L_t^* and C_t^* to denote teachers' optimal choices for learning and care activities. By differentiating these FOCs 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:

$$0 = zu_{HH} f_1 \left(f_1 \frac{dL_t^*}{dA} + f_2 \frac{dC_t^*}{dA} + f_3 \right) + zu_H \left(f_{11} \frac{dL_t^*}{dA} + f_{12} \frac{dC_t^*}{dA} + f_{13} \right) + u_{KK} \frac{dL}{dA} + u_{KK} \frac{dC_t^*}{dA} \tag{D.4}$$

$$0 = zu_{HH} f_2 \left(f_1 \frac{dL_t^*}{dA} + f_2 \frac{dC_t^*}{dA} + f_3 \right) + zu_H \left(f_{12} \frac{dL_t^*}{dA} + f_{22} \frac{dC_t^*}{dA} + f_{23} \right) + u_{KK} \frac{dL_t^*}{dA} + u_{KK} \frac{dC_t^*}{dA} \tag{D.5}$$

using the fact that at the optimum we have $f_1 = f_2 = u_K/(zu_H)$, we get that:

$$0 = zu_{HH} \frac{u_K}{zu_H} \left(\frac{u_K}{zu_H} \frac{dL_t^*}{dA} + \frac{u_K}{zu_H} \frac{dC_t^*}{dA} + f_3 \right) + zu_H \left(f_{11} \frac{dL_t^*}{dA} + f_{12} \frac{dC_t^*}{dA} + f_{13} \right) + u_{KK} \frac{dL}{dA} + u_{KK} \frac{dC_t^*}{dA} \tag{D.6}$$

$$0 = zu_{HH} \frac{u_K}{zu_H} \left(\frac{u_K}{zu_H} \frac{dL_t^*}{dA} + \frac{u_K}{zu_H} \frac{dC_t^*}{dA} + f_3 \right) + zu_H \left(f_{12} \frac{dL_t^*}{dA} + f_{22} \frac{dC_t^*}{dA} + f_{23} \right) + u_{KK} \frac{dL_t^*}{dA} + u_{KK} \frac{dC_t^*}{dA} \tag{D.7}$$

Combining these expressions yields:

$$\frac{dC_t^*}{dA} = \frac{f_{12} - f_{11}}{f_{12} - f_{22}} \frac{dL_t^*}{dA} + \frac{f_{23} - f_{13}}{f_{12} - f_{22}} \tag{D.8}$$

which, substituting back into equation (D.6) and equation (D.7) gives equations (6.5) and (6.5) from the main text:

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

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

where

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

D.2 Specialized Model with Misperceptions

In this Section we show more details of the specialization of this general model of teachers' behavior, covered in Section 6.3 in the main text. In particular, we consider two steps in the teachers' decision process. This approach is legitimate if the (dis) utility teachers get from L_t and C_t is the same, which is implicit in our specification of the utility function $u(H, 1 - L_t - C_t)$. Given the total amount of time in the classroom, therefore, teachers maximise H. Given the optimal decision making in the second stage, they decide on the allocation between N and K. We begin by discussing their second-stage problem and then briefly outline their first-stage problem.

D.2.1 Teachers' second-stage problem: allocating time between learning and care

Conditional on the total amount of TA time A and the total amount of teacher classroom time N, teachers' split their own time and their TA's time so as to maximize child development. Let τ_t be the fraction of teacher time spent on learning, and τ_a be the fraction of TA time spent on learning. The teacher's second-stage problem is thus:

$$\max_{\tau_a, \tau_t} \tilde{z}\big(\tilde{w}^{(1-\rho)}\tilde{L}^{\rho} + (1-\tilde{w})^{(1-\rho)}\tilde{C}^{\rho}\big)^{\frac{1}{\rho}} \tag{D.11}$$

$$s.t. \tag{D.12}$$

$$\tilde{L} = \big(\tilde{\theta}_l \tau_t^{\lambda} N^{\lambda} + (1-\tilde{\theta}_l)\tau_a^{\lambda} A^{\lambda}\big)^{\frac{1}{\lambda}}$$

$$\tilde{C} = \big(\tilde{\theta}_c (1-\tau_t)^{\lambda} N^{\lambda} + (1-\tilde{\theta}_c)(1-\tau_a)^{\lambda} A^{\lambda}\big)^{\frac{1}{\lambda}} \quad \lambda \in (0,1] \tag{D.13}$$

The corresponding first order conditions, which are given in the main text, are:

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

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

where $h = \big(\tilde{w}^{1-\rho}\tilde{L}^{\rho} + (1-\tilde{w})^{1-\rho}\tilde{C}^{\rho}\big)$. 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 (6.10) the expressions for $\partial\tilde{L}/\partial\tau_t$ and $\partial\tilde{C}/\partial\tau_t$, we get equation (6.12):

$$\Big(\frac{\tilde{w}}{1-\tilde{w}}\Big)^{1-\rho}\frac{\tilde{\theta}_l}{\tilde{\theta}_c} = \Big(\frac{h_c}{h_l}\Big)^{\frac{\rho}{\lambda}-1}\Big(\frac{\tau_t}{1-\tau_t}\Big)^{1-\lambda} \tag{D.16}$$

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 (6.11), we obtain:

$$\Big(\frac{\tilde{w}}{1-\tilde{w}}\Big)^{1-\rho}\frac{1-\tilde{\theta}_l}{1-\tilde{\theta}_c} = \Big(\frac{h_c}{h_l}\Big)^{\frac{\rho}{\lambda}-1}\Big(\frac{\tau_a}{1-\tau_a}\Big)^{1-\lambda} \tag{D.17}$$

Taking the ratio of these two equations we obtain equation (6.15) in the main text:

$$\frac{\tilde{\theta}_l}{1-\tilde{\theta}_l}\frac{1-\tilde{\theta}_c}{\tilde{\theta}_c} = \Big(\frac{\tau_t}{\tau_a}\Big)^{1-\lambda}\Big(\frac{1-\tau_a}{1-\tau_t}\Big)^{1-\lambda} \tag{D.18}$$

And when there are no TAs we get:

$$\frac{\tilde{w}}{1-\tilde{w}} = \frac{\tau_t}{1-\tau_t} \tag{D.19}$$

D.2.2 A change in $\tilde{w}$, the perceived importance on learning *vs.* care activities

Taking logs of expression (D.18) gives:

$$\log(\tilde{\theta}_l)-\log(1-\tilde{\theta}_l)+\log(1-\tilde{\theta}_c)-\log(\tilde{\theta}_c) = (1-\lambda)\log(\tau_t)-(1-\lambda)\log(\tau_a)+(1-\lambda)\log(1-\tau_a)-(1-\lambda)\log(1-\tau_t) \tag{D.20}$$

Holding fixed N and totally differentiating equation (D.20) with respect to $\tilde{w}$, gives:

$$\frac{d\tau_a}{dw} = \frac{\tau_a(1-\tau_a)}{\tau_t(1-\tau_t)}\frac{d\tau_t}{dw} \tag{D.21}$$

This shows that the any change in teachers' perception of the relative importance of learning vs. care routines will, holding fixed their total time input, lead to a proportional change in both the fraction of time they allocate to learning and the fraction of time their TA does. Both changes will always be of the same sign.

Taking logs of equation (D.16) gives:

$$(1-\rho)\log(\tilde{w}) - (1-\rho)\log(1-\tilde{w}) + \log(\tilde{\theta}_l) - \log(\tilde{\theta}_c) = \frac{\lambda-\rho}{\lambda}\log h_l - \frac{\lambda-\rho}{\lambda}\log h_c + (1-\lambda)\log\tau_t - (1-\lambda)\log(1-\tau_t) \tag{D.22}$$

Holding N fixed and totally differentiating gives with respect to $\tilde{w}$ gives:

$$\frac{1-\rho}{\tilde{w}(1-\tilde{w})} = \frac{\lambda-\rho}{\lambda}\frac{1}{h_l}\frac{dh_l}{dw} - \frac{\lambda-\rho}{\lambda}\frac{1}{h_c}\frac{dh_c}{d\tilde{w}} + \frac{1-\lambda}{\tau_t(1-\tau_t)}\frac{d\tau_t}{d\tilde{w}} \tag{D.23}$$

where

$$\begin{aligned}\frac{dh_l}{d\tilde{w}} &= \tilde{\theta}_l N^\lambda \lambda \tau_t^{\lambda-1}\frac{d\tau_t}{d\tilde{w}} + (1-\tilde{\theta}_l)A^\lambda\lambda\tau_a^{\lambda-1}\frac{d\tau_a}{d\tilde{w}} \\ &= \frac{d\tau_t}{d\tilde{w}}\left[\tilde{\theta}_l N^\lambda\lambda\tau_t^{\lambda-1} + (1-\tilde{\theta}_l)A^\lambda\lambda\tau_a^{\lambda-1}\frac{\tau_a(1-\tau_a)}{\tau_t(1-\tau_t)}\right] \qquad \text{(D.24)} \\ \frac{dh_c}{dw} &= -\tilde{\theta}_c N^\lambda\lambda(1-\tau_t)^{\lambda-1}\frac{d\tau_t}{dw} - (1-\tilde{\theta}_c)A^\lambda\lambda(1-\tau_a)^{\lambda-1}\frac{d\tau_a}{dw} \\ &= -\frac{d\tau_t}{dw}\left[\tilde{\theta}_c N^\lambda\lambda(1-\tau_t)^{\lambda-1} - (1-\tilde{\theta}_c)A^\lambda\lambda(1-\tau_a)^{\lambda-1}\frac{\tau_a(1-\tau_a)}{\tau_t(1-\tau_t)}\right] \qquad \text{(D.25)}\end{aligned}$$

Combining, we have:

$$\frac{d\tau_t}{d\tilde{w}} = \frac{\frac{1-\rho}{\tilde{w}(1-\tilde{w})}}{\frac{1-\lambda}{\tau_t(1-\tau_t)} + \frac{\lambda-\rho}{\lambda}\frac{1}{h_l}\left[\tilde{\theta}_l N^\lambda \lambda \tau_t^{\lambda-1} + (1-\tilde{\theta}_l)A^\lambda \lambda \tau_a^{\lambda-1}\frac{\tau_a(1-\tau_a)}{\tau_t(1-\tau_t)}\right] + \frac{\lambda-\rho}{\lambda}\frac{1}{h_c}\left[\tilde{\theta}_c N^\lambda \lambda (1-\tau_t)^{\lambda-1} - (1-\tilde{\theta}_c)A^\lambda \lambda (1-\tau_a)^{\lambda-1}\frac{\tau_a(1-\tau_a)}{\tau_t(1-\tau_t)}\right]}$$

$$> 0 \tag{D.26}$$

And, by expression (D.21), we have $\frac{d\tau_c}{dw} > 0$

D.2.3 First stage: allocating time between leisure and the classroom

Taking the production stage as given, teachers choose how much time to allocate to classroom activities:

$$\max_N \quad u(H(N,A), 1-N) \tag{D.27}$$

FOC:

$$0 = u_H \frac{dH}{dN} - u_K(1-N)$$