IDB WORKING PAPER SERIES Nº IDB-WP-951

Quality of Parenting Programs and Child Development Outcomes:

The Case of Peru's Cuna Mas

M. Caridad Araujo
Marta Dormal
Marta Rubio-Codina

Inter-American Development Bank
Social Protection and Health

October 2018

Inter-American Development Bank

Quality of Parenting Programs and Child Development Outcomes:

The Case of Peru's Cuna Mas

M. Caridad Araujo
Marta Dormal
Marta Rubio-Codina

October 2018

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Araujo, Maria Caridad.
Quality of parenting programs and child development outcomes: the case of Peru's Cuna Mas / M. Caridad Araujo, Marta Dormal, Marta Rubio-Codina.
p. cm. — (IDB Working Paper Series ; 951)
Includes bibliographic references.
1. Home-based family services-Peru. 2. Parenting-Government policy-Peru. 3. Child care services-Government policy-Peru. 4. Child development-Government policy-Peru. 5. Child welfare-Government policy-Peru. I. Dormal, Marta. II. Rubio-Codina, Marta. III. Banco Interamericano de Desarrollo. Social Protection and Health Division. IV. Title. V. Series.
IDB-WP-951

http://www.iadb.org

scl-sph@iadb.org

www.iadb.org/SocialProtection

Quality of Parenting Programs and Child Development Outcomes: the Case of Peru's *Cuna Mas*

M. Caridad Araujo[i], Marta Dormal[i], Marta Rubio-Codina[i]

Abstract

This paper analyzes the association between quality of parenting programs and child development outcomes. It focuses on parenting programs delivered through home visits (i.e., home visiting programs) in which a paraprofessional works with the caregiver to improve parental practices, skills and knowledge of child development, with an application to the Peruvian national program *Programa Nacional Cuna Más.* Home visiting quality is measured through the *Home Visit Rating Scale, version A+2* (HOVRS-A+2), and child development outcomes through the *Ages and Stages Questionnaires* (ASQ-3) and the *Bayley Scales of Infant and Toddler Development* (Bayley-III). The empirical analysis used is descriptive in nature. Results show that home visiting quality scores are significantly associated with our measures of child development, but this association is not consistent across all measures.

JEL codes: I12, J24, O15

Keywords: home visiting, quality, child development, HOVRS-A+2, ASQ-3, Bayley-III

[i] Social Protection and Health Division, Inter-American Development Bank, Washington DC, USA.

1. Introduction

Parenting interventions, and particularly those delivered through home visits, have gained attention in policy circles in Latin America and the Caribbean (LAC) and in other regions of the world in recent years. Parenting programs consist of interventions aimed at changing behaviors and improving parental practices, skills and knowledge of child development, through training, support and education from a paraprofessional working directly with the child's caregiver. Three delivery models are common: group sessions, home visits or clinic appointments (Berlinski and Schady 2015). Informed by evidence—mostly coming from efficacy trials and small-scale, rigorously evaluated pilots, and that has shown that these programs can have high and sustained impacts on child development (Britto et al. 2016)—countries like Peru, Brazil, and others have undertaken the challenge of expanding the coverage of the home visiting model ("home visiting program" hereinafter) amongst disadvantaged families with young children.

As with any social service, scaling-up raises important implementation challenges. In home visiting, given it is an area relatively less researched, little is known about the attributes of service quality that are most critical to preserve during the process of expansion. Is it enough to warrant a minimum dosage and length of visits? What competencies ought to be mastered by home visitors to successfully modify parental behaviors and promote high-quality interactions and appropriate psychosocial stimulation in the home? And more broadly, what are the key components of a successful home visit for ensuring sustained impacts on child development outcomes at scale?

Paulsell et al. (2010) propose three dimensions of home visiting quality: dosage, content, and relationships. Dosage and content can be understood as *structural* elements of quality, while relationships are commonly referred to as *process* quality characteristics. Process quality focuses on the dynamics aspects of the visit and on how the intervention content is delivered, particularly regarding relationships and the quantity and nature of the interactions that occur between the home visitor, the caregiver, and the child. Structural quality, on the other hand, refers to the presence or absence of resources that facilitate these interactions during the home visit (e.g., materials, resources, home visitor's experience, education and salary). These two dimensions of quality are highly interrelated. For instance, the home visitor is much more likely to engage in high-quality interactions with the child and caregiver when she is provided with appropriate learning materials to support her work. Their measurement, however, differs considerably: process quality requires expert observation, judgment and interpretation of these interactions and is therefore more complicated and lengthy to measure than structural quality, which is easily quantifiable and can be evaluated through a variety of checklist-type tools.

Several instruments have been used to measure home visiting quality. By and large, these instruments place emphasis on measuring the overall quality of the home visit and of the home visitor's relationship with the caregiver. Instruments that have been

used in published studies are:

- The *Home Visit Rating Scale series* (HOVRS, HOVRS-A, HOVRS-A+ and HOVRS-A+2; Roggman et al. 2006, Roggman et al. 2014). It consists of seven scales organized in two wider domains, the first focusing on the strategies used by the home visitor to effectively carry out the visit, and the second measuring the level of engagement of caregiver and child and the nature of their interactions.
- The *Home Visit Content and Characteristics Form* (HVCCF; Boller et al. 2009). Sometimes used as a complement of the HOVRS, it documents a range of structural characteristics of the home visit, such as dosage and content.
- The *COACH rating system* (Dishion et al. 2010), which is an observational system designed to quantify fidelity of content delivery, with a focus on five dimensions: conceptual accuracy, observance of and responsiveness to family's needs, structure of the visit, careful and appropriate teaching, and motivation.

A common element to the administration and scoring of all these instruments is the requirement of rigorously trained observers, vastly knowledgeable of the scoring protocol of each one of them.

Schodt et al. (2015) offers a review of the literature on measuring quality of home visiting and of the few studies exploring the association between home visiting quality and child development outcomes. These studies, briefly discussed next, are broadly descriptive: they present correlations between quality of home visits, the quality of the environment in the home, and child development. Nonetheless, they lack a research design that can establish a causal relationship between high-quality visits and child development.

Vogel et al. (2015) is part of the Baby FACES study, an ongoing longitudinal descriptive evaluation of the US Early Head Start program. The authors characterized home visiting quality using the HOVRS-A and the HVCCF and explored its association with child development outcomes at age two. They documented a positive significant association between the second domain of HOVRS-A, Home visitor effectiveness, and the Child engagement scale, and the child's emotional regulation score, measured by the Bayley Behavior Rating Scale. The order of magnitude of these associations ranged between 0.13 and 0.14. Consistently, they also reported a negative association of -0.17 between home visiting quality and home-visitor reported child behavioral problems, as assessed by the Brief Infant Toddler Social Emotional Assessment. Nonetheless, they documented negative and significant correlations between home visit quality and gross motor, fine motor and problem solving, as measured using the ASQ-3. The size of these associations ranged between -0.14 and -0.17. No significant correlation was found between measures of home visiting quality and quality of the home environment as measured by the HOME.

Smith et al. (2013) analyzed data from the Family Check-Up program, also in the US. This model provides parents with tools to manage children's behavioral problems. The

authors used the COACH rating system to characterize the fidelity of content delivery and explored the indirect effect of home visiting quality on child outcomes. Specifically, they found that home visitors who were more competent in adhering to program content were better able to engage parents in the session, and this translated into a small but significant improvement in parental abilities in the use of the techniques of positive behavior support promoted by the Program (0.06 of a standard deviation, SD). One year later, their children also exhibited a reduction in problem behaviors as measured by the externalizing scale of the Child Behavior Checklist (-0.24 SD).

Roggman et al. (2016) examined data on home visiting quality from two Head Start program sites in the United States. They showed a significant association between home visiting quality measured by the HOVRS and the quality of the home environment measured by the HOME. The partial correlations of the HOVRS domains and total score with the HOME, controlling for program site, ranged between 0.31 and 0.34. Using regression analysis, they explored the association between home visiting quality and child receptive vocabulary outcomes measured by the PPVT at 36 months of age and found a positive and significant association, controlling for program site. This coefficient was no longer significant once the quality of the home environment was accounted for in the regression, suggesting that the impact of the HOVRS on child outcomes could be operating indirectly through parenting practices.

In this paper, we explore the association between quality of home visiting and child development outcomes in the context of a program implemented in a middle-income country. We focus on the Peruvian home visiting service delivered by the *Programa Nacional Cuna Más*, known as the *Servicio de Acompañamiento a Familias* (Cuna Mas, hereafter). We measure quality using the *Home Visit Rating Scale, version A+2* (HOVRS-A+2). Child development outcomes are measured using two different instruments: the *Ages and Stages Questionnaires* (third version, ASQ-3; Squires et al. 2009) and the *Bayley Scales of Infant and Toddler Development* (third version, Bayley-III; Bayley 2006). This analysis builds on the impact evaluation of Cuna Mas by providing an in-depth look at quality of implementation in a subsample of families and home visitors. The impact evaluation of Cuna Mas showed robust and significant impacts on child development, in the areas of cognitive development and communication in the order of magnitude of 0.14 to 0.17 SD (Araujo et al. 2018). The theory of change behind home visiting programs is that home visitors support caregivers in developing parental practices which create a favourable learning environment for the child in the home, which in turn translate into better child developmental outcomes. In line with this theory of change, the evaluation of Cuna Mas suggested that impacts on child development seemed to be mediated through several parental behaviors, such as more frequent play and learning activities in the home, more positive discipline and larger enrolment into preschool (MEF-BID 2016). The hypothesis behind the present study is that the *quality* of service delivery—and particularly process quality—matters to improve the above-mentioned parental practices and child developmental outcomes. In other words, we would expect home visiting programs delivered with higher quality, through better home visitor-

caregiver/child interactions, to be more strongly associated with improvements in parental behaviour changes, and consequently with child development outcomes. The empirical analysis herein is descriptive in nature and focuses on the associations between child development outcomes and quality measures.

The remainder of this paper is organized as follows. Section 2 briefly describes the Cuna Mas program and Section 3 the study design and the data. Section 4 presents summary statistics of the analysis sample and Section 5 reports the associations between quality of home visits and child development. Finally, Section 6 concludes.

2. The Cuna Mas home visiting program

Cuna Mas is Peru's largest national early childhood development service provider, reaching over 93 thousand children 0-36 months of age in rural areas in 2015 with its home visiting services. It is a public program that is fully subsidized. It delivers one-hour weekly home visits, carried out by a community member who has been trained by the Program. Although the Program requires home visitors to meet certain minimum requirements—being literate, 21 years of age or older, speaking the local language, having experience working with children, and having a certain level of recognition within the community—these are often difficult to find in the context of the poor, disperse rural communities where the Program operates. Community Council members[1] nominate a list of candidates to the Program and local supervisors select home visitors from that pool of applicants. Every home visitor oversees ten families. In turn, for every ten home visitors, there is one supervisor who is expected to work closely helping them plan their work and providing unstructured on-the-job training and mentoring. In practice, however, supervisors are often tasked with other time-consuming administrative tasks.

The home visitor follows a structured curriculum that promotes more and better-quality adult-child interactions and learning through play. When this study was carried out, each visit was organized around three 'moments'. The first moment, 'Family Life' (*Vida en Familia*), seeks to help caregivers identify opportunities for learning and play in their daily routine activities. Often, more general messages related to health, nutrition, and sanitation are also delivered. The second moment, 'Learning through Play' (*Jugando Aprendo*), uses intentional play activities and materials—i.e., toys, and form boards, either provided by the Program or homemade with the caregiver—to promote the development of language, fine motor and cognitive skills. The third moment, 'Tell me a Story' (*Cuéntame un Cuento*) encourages vocabulary acquisition and language development using picture books provided by the Program, and helps caregivers learn how to use these materials with infants and toddlers. In addition to these moments, at the beginning of each visit, the home visitor and the caregiver spend time revising the activities form the previous visit[2]. Similarly, at the end of each visit, they recap and identify the specific activities that caregiver and child will practice during the coming

[1] *Comités de Gestión*, as they are known in the areas.
[2] Starting in 2017, the Program started revising its intervention model and contents to emphasize the nutritional components over the child development ones.

week. To this end, all materials used in the visit are left in the home and collected in the next visit. The home visitor rotates these materials amongst the families she works with, as a 'toy library'.

Cuna Mas targets the poorest in the country and, in particular, districts[3]: (i) with a poverty rate of at least 50 percent; (ii) with a chronic malnutrition rate of at least 30 percent; (iii) targeted by the conditional cash transfer program *Juntos*; and (iv) where over 50 percent of the population lives in rural communities (*Centro Poblados* or CP). Within these districts, Cuna Mas intervenes exclusively in rural CPs, which are defined as those with less than 2,000 families or 400 dwellings. All children from birth to 24 months in these communities are eligible to enroll in the Program and receive visits until they are 36 months of age, when they graduate.

An experimental evaluation of Cuna Mas was built taking advantage of its gradual roll out to assess its impact on child development outcomes. The impact evaluation sample included 5,859 children ages 1-24 months at baseline in 360 CPs in 180 eligible districts. These districts were spread across 12 departments out of the 24 in the country, mostly located in the Andean and Amazon regions. 120 of these districts were randomly assigned to the treatment group and received visits, and 60 districts were assigned to the control group and did not receive the intervention. Baseline data was collected in mid-to-late 2013 and endline data was collected approximately 24 months later. DGSE-MIDIS et al. (2015) offers more details on the evaluation sample and design and Araujo et al. 2018 reports the evaluation findings.

3. Study design, sample and data

3.1. Quality data sample

Cuna Mas implementation began in August 2013 in all treatment districts of the impact evaluation sample. As shown in Figure 1 (Panel A), endline data collection was carried out between June and December 2015, where two developmental outcomes for children were assessed (the ASQ-3 and the Bayley-III). In parallel, between August and October 2015, home visit quality data was collected for a subsample of those children. The Program had been operating for approximately two years when quality was measured.

This subsample was selected from the endline impact evaluation sample, as follows: we first identified eligible children as those receiving visits at the time of data collection—i.e., children in the treatment group and younger than 36 months; and then, we identified their home visitor as well as three or four additional children visited by her but who were not necessarily part of the evaluation sample. The quality data sample ended up including information on 554 home visits carried out to an equal number of children, of which 229 were focus children (i.e., also part of the evaluation sample). Participants lived in 103 rural communities and were visited by 176 home visitors.

[3] Districts are the equivalent of municipalities, or counties, in rural Peru.

3.2. The data: measures, adaptation and training

3.2.1. Home visit quality: the HOVRS-A+2

For each child in the sample, we filmed one home visit and rated various aspects of its quality and content using suitably translated and adapted versions of the *Home Visit Rating Scale, version A+2* (HOVRS-A+2; Roggman et al. 2014).

Developed from the original HOVRS (Roggman et al. 2006), the HOVRS-A+2 is the latest of a series of instruments specifically designed to assess home visit process quality. It consists of seven scales organized in two wider domains. The first domain, *Home Visitor Practices*, includes four scales that focus on the strategies used by the home visitor to effectively carry out the visit—this is to say, the extent to which the home visitor is prepared for the visit, observes and responds to caregiver and child, and shows a warm and positive attitude towards them, amongst others. The second domain, *Family Engagement,* includes three scales aimed at measuring the nature of the interactions between caregiver and child, as well as their interest and level of participation throughout the visit.[4] It is administered by observation and was initially developed for scoring a video recording of the visit, but is now used also for direct scoring during a live visit (L. Roggman, personal communication, November 16, 2017).

The HOVRS family of scales has been widely used in the US to measure home visit quality, both in the context of research studies and to support programmatic efforts. It has proven adaptable to diverse program goals, visit formats, and different cultures within the US, such as Spanish-only Latino families, rural Caucasian families, and urban African-American families.[5]

Adaptation of the HOVRS-A+2 and training of the HOVRS-A+2 supervisor[6]. The HOVRS-A+2 was translated to Spanish and back-translated into English to ensure linguistic and functional equivalence. The back-translation was reviewed by the developers, who suggested minor modifications. The developers trained the person who supervised the video coding phase and who led the training of the actual coders. The HOVRS-A+2 supervisor´s training lasted for two days, following which she continued to score videos of home visits until satisfactory inter-observer reliabilities were obtained.

[4] The scales in the *Home Visitor Practices* domain are: (i) Home visitor responsiveness to family (6 items); (ii) Home visitor relationship with family (8 items); (iii) Home visitor facilitation of parent–child interaction (6 items); and (iv) Home visitor non-intrusiveness and collaboration (5 items). Those in the *Family Engagement* domain are: (i) Parent-child interaction during the home visit (7 items); Parent engagement during the home visit (6 items); and (iii) Child engagement during the home visit (4 items).

[5] For more information, see: http://healthandwelfare.idaho.gov/Portals/0/Children/HomeVisiting/MIECHV%20Assessment%20Guide%204-9-2013.pdf

[6] We refer to “HOVRS-A+2 supervisor" as the person who trained the HOVRS coders and supervised them throughout the coding process.

Training and filming of videos. The home visits were filmed between August and October 2015 by a team of 14 individuals, who had been trained for a week, including practices. A total of 24 people participated in the training, half of which had prior experience filming. The field team was selected based on their performance during the training and practices. The content of the training included how to use a video camera, how to select the optimal angle for filming, and other strategies to obtain good quality videos given the limitations of the setting (i.e., filming indoors, background noise, static camera, etc.).

The training also introduced the Program, the structure of a home visit, and revised the filming protocol designed to minimize the disturbance created to the family and to the visit itself. This was done to prevent video-recording from affecting the behavior of participants during the visit. Supervisors were instructed to discuss with each family and home visitor the objective of this exercise prior to the start of the data collection. There were no reports of disruptions or conflicts during data collection.

Training of coders. The coding of the videos was carried out between November 2015 and May 2016 by a team based in Quito, Ecuador, with experience in the coding of observational instruments to measure quality on videos of child care, preschool, and elementary school settings. Nine coders and a supervisor (a psychologist) were trained for six days. The training included between 6-8 practice codings, until acceptable reliability was consistently obtained. The five coders with the highest reliability with the HOVRS-A+2 supervisor—all with higher education studies in psychology, education or related fields—were selected to code the study videos. There was also a team coordinator, a psychologist, who acted as a back-up coder.

Two scores on the HOVRS-A+2 were considered non-reliable if: (i) one or more scales differed by more than one point (scale-level reliability criterion); or (ii) within a given scale, two or more items differed by two or more points (item-level reliability criterion).[7] Reliability between the coders and the HOVRS-A+2 supervisor was acceptable: throughout the training, item-level reliability ranged from 80 to 100 percent on average per coder and was above 90 percent for three of the coders; and scale-level reliability ranged from 61 to 86 percent on average per coder, with values above 80 percent for three coders.

Coding process. Each video was coded twice, independently, by two coders randomly selected from the team. If scores from both coders did not satisfy the item-level or scale-level reliability criteria, the video was assigned to a third coder—different from the first two and also selected at random—for a third coding.

To sustain reliability throughout the coding period, we put in place a calibration exercise whereby the team coded a 20-minute fragment daily. Scores were compared to the HOVRS-A+2 supervisor's and disagreements discussed. Half way through the training, however, we modified the calibration exercise to the coding of an entire visit

[7] Items in the HOVRSA-A+2 are scored on a 4-point scale, with possible values 1, 3, 5 or 7. As such, a 1-point difference at the item-level corresponds to, for example, the difference between 1 and 3, 3 and 5, or 5 and 7.

twice a week since it was felt that coding 20-minute fragments generated fewer disagreements than coding 1-hour visits.[8] This strategy significantly reduced the number of videos that had to be coded a third time. The percentage agreement of the coders with the HOVRS-A+2 supervisor during this exercise was very high, ranging from 93 to 98 percent.

Of the 554 videos in the quality sample, a total of 214 videos (39 percent) were coded a third time: 142 videos (66 percent) did not comply with the scale-level reliability criterion, 17 (8 percent) did not comply with the item-level reliability criterion, and 55 videos (26 percent) did not comply with either criterion. After a third coding, average reliability increased from 65 to 82 percent at the scale level and from 87 to 94 percent at the item level.[9]

The average video was about one-hour long. The team coded no more than three videos a day in four-hour sessions, which included the calibration exercise. Any verbal interactions in the video that were in Quechua were subtitled in Spanish by a team of Peruvian translators prior to coding. This occurred in 13 percent (74) of the videos.

Scoring. Each *item* describes a specific practice and is scored on a 4-point scale, with anchor points of 1 'needs training or support', 3 'adequate', 5 'good', and 7 'excellent'. The coder then assigns a score to the *scale* from 1 to 7 based on the general pattern of scores observed for the items.[10] We categorize scores following the developers' suggested groups: 'needs training or support' (if 1 ≤average scale score <3), 'adequate' (if 3 ≤average scale score <5), 'good' (if 5 ≤average scale score <7) and 'excellent' (if average score = 7). Domain scores are constructed averaging scale scores within a domain and a total test score is constructed as the average of the two domains.

In this paper, we work with the domain and total scores. For each visit, we keep the scores of the two coders with the highest scale- and item-level reliability. Although coders were selected at random and strict reliability criteria were maintained throughout the coding process, idiosyncrasies across coders in the coding process might remain. We therefore ran a regression on coder dummies to remove the coders' fixed-effect. We then averaged the residuals of these regressions for the two more reliable coders and normalized the resulting score to have mean zero and SD of one.

3.2.2. Child development outcomes: the ASQ-3 and the Bayley-III

We measured child development outcomes using the *Ages and Stages Questionnaires* (third version, ASQ-3; Squires et al. 2009) and the *Bayley Scales of*

[8] Interactions of different nature are more likely to occur during the full hour than in 20 minutes.
[9] For this exercise, average reliability at the scale level is constructed by first computing the percentage of scales that meet the scale-level reliability criterion for each video and then taking the sample average. An analogous procedure was used for the average reliability at the item level.
[10] The developers recommend that scale-level scores are assigned on a scale from 1 to 7 based on the general pattern of scores observed for the items This results in very similar scores to those obtained by computing the mathematical average of all items within the scale (correlations range from 0.92 for *Responsiveness to family* to 0.97 for *Child engagement* during home visit).

Infant and Toddler Development (third version, Bayley-III; Bayley 2006). The ASQ-3 was administered to all children in the sample at two points in time: first, at baseline, between April and August 2013, when they were 1-24 months old and before they started receiving home visits; and a second time, at endline, between June and December 2015, when they were 25-55 months and after approximately two years of receiving home visits. The Bayley-III was administered only at endline, between mid-July and mid-December 2015, and only to the sub-sample of children who at that time were younger than 42 months (the test is not designed to assess older children).

The ASQ-3 is a screening tool, comprised of age-specific caregiver-completed questionnaires. Each questionnaire assesses child competence in five domains—problem solving (or cognition), communication, fine motor, gross motor, and personal-social—with six items in each. Because of the low education levels of some caregivers in the sample, items were given by interview or were directly administered to the child by the interviewer. A team of Peruvian psychologists adapted the Spanish version of the test to the context and determined which items were to be administered by caregiver report and which by direct administration.

The ASQ-3 has become increasingly used in large-scale impact evaluation studies (Ángeles et al. 2012; Fernald et al. 2012; Bernal 2015). As a screener, it is designed to identify children at risk of developmental delay and therefore is more sensitive for the measurement of development at the lower end of the distribution of skills. Since our aim was to identify changes in developmental outcomes due to the intervention across the entire distribution of developmental ability, we modified the administration and gave the first three new items from subsequent questionnaires whenever the child attained the maximum score in a scale. If test items are arranged in increasing order of difficulty, this extends the ceiling of the test. Similar adaptations of the ASQ have been used in other studies in middle- or low-income countries (Fernald et al. 2012).

We also administered the Bayley-III to the sub-sample of children for which the test is available. A diagnostic test, the Bayley-III is widely recognised as the 'gold standard' to measure the developmental levels of children from birth to 42 months of age (Frongillo et al. 2014; Fernald et al. 2012; Fernandes et al. 2014). Moreover, it has shown sensitivity to the effects of similar programs to Cuna Mas in contexts as diverse as Bangladesh (Nahar et al. 2012) and Colombia (Attanasio et al. 2014).

The Bayley-III assesses five developmental domains: cognition, receptive and expressive language, fine and gross motor, socio-emotional and adaptive behaviour. Due to budget and time constraints, we only administered those scales more directly linked to the intervention aims—namely, cognition, receptive and expressive language, and fine motor. These scales are entirely assessed by direct observation of the child's performance in a series of tester-administered items, arranged in increasing order of difficulty; and basal and ceiling rules determine the number of items to administer to each child.

The test was translated and back-translated to ensure linguistic and functional

equivalence. In addition, various items required adaptation and contextualization, which were finalised following extensive piloting. The process was led by a Peruvian psychologist, who was also involved in the adaptation of the ASQ-3 for this study (Lazarte 2016 provides details).

Training and preparation for testing. The ASQ-3 was administered at the child's home, following a household survey. The baseline survey was carried out by a team of 76 interviewers, who were trained by 3 trainers for 14 days, including 5 days of practices. At endline, to improve data quality, the team of interviewers in charge of the survey was reduced to 20, with a similarly long training (13 days) and more practice administrations. For the sample where home visiting quality was measured—a subsample of the impact evaluation sample—a total of 52 interviewers collected the ASQ-3 at baseline and 18 at endline.

The Bayley-III requires the test to be administered by child development professionals, such as psychologists and educators, in a quiet room that facilitates the child's concentration on the assessment. Twelve psychology graduates were trained on the Bayley-III for 5.5 weeks, including 15-18 practices each. Of them, 8 were selected to administer the test and 2 were assigned a supervisory role. All children in the sample were assessed in a room adapted for this purpose within their village.

Practice testing occurred in pairs and continued until trainee-trainer agreement was >0.9 on each test. During field activities, 7.55 percent of the Bayley-III assessments were observed by the trainer and corrective feedback was given when appropriate. The agreement between tester and trainer scores during these tests was over >0.98, ensuring high data quality throughout. All assessments took place in the presence of the child's main caregiver (often the mother) after written consent was obtained.

Obtaining a child development score. For each item in the ASQ-3, answer options 'yes', 'sometimes' or 'not yet' are allocated 10, 5 or 0 points, respectively. For each scale, the total raw score is the sum of all items within the scale. Missing items are replaced with the scale average, although if more than two items are missing in a scale, then the scale is not scored.

On the Bayley-III, the child scores 1 for each item correctly executed and 0 otherwise. For each scale, the raw score is the sum of correct responses, including non-administered items preceding the basal. By construction, the Bayley-III raw scores increase with age.

Since the ASQ-3 has age-specific questionnaires, raw scores should be constant with age. However, in practice, they exhibited a pattern with age. Therefore, we first transform the raw scores for each domain in the ASQ-3 or Bayley-III to z-scores after non-parametrically adjusting for child age using the method proposed by Rubio-Codina et al. (2016a). Unlike using norms from the reference populations for each test, this standardization method has the advantage that it handles age effects consistently across tests. We then constructed different aggregate scores, as follows. For the ASQ-3, we construct three aggregates: in one, the five domains of the test

(cognition, language, fine motor, gross motor, and personal-social) each receive the same weight; a second, narrower aggregate uses only the cognition, language, and fine motor domains, with each receiving the same weight; and a third aggregate only uses the cognition and language domains. For the Bayley-III, we proceed in an analogous fashion, and also constructed three aggregates. The first includes the four domains (cognition, receptive language, expressive language, and fine motor), with each receiving the same weight; in the second aggregate, we first construct a total language score as the average of the receptive and expressive language scores and then, in a second step, aggregate the cognition, language, and fine motor domains, with each receiving one-third of the total weight. We construct this aggregate in this way because it most closely mimics the second aggregate using the ASQ-3 data; finally, as with the ASQ-3, we construct an aggregate that is only based on the cognition and language domains. All aggregates are then re-standardized so that they have zero mean and unit standard deviation.

3.2.3. Other measures

A household survey was collected both at baseline and at endline. It included basic socio-economic information, the child's health history and maternal knowledge on child development, amongst other variables. The survey additionally collected the following indices of the quality of the home environment: the acceptance and responsivity scales from the *Home Observation for Measurement of the Environment* (HOME, Caldwell and Bradley 2003), as well as the number of varieties of play materials the child usually played with, and the number of varieties of play activities the child and an adult in the home engaged in over the week prior to the survey using the *Family Care Indicators* (FCI, Frongillo et al. 2003). Play materials were recorded by direct observation and play activities by caregiver report.[11] The survey also measured parental knowledge of child development through the *Knowledge of Infant Development Inventory* (KIDI, MacPhee 2002). All indices were constructed by factor analysis and normalized to have a mean of zero and SD of one.

At endline only, a home visitor and a community survey were also collected. The home visitor questionnaire comprised socio-economic and other relevant information, such as prior experience, time on the job, and knowledge on child development. The community survey collected data on the availability of services and distances to larger towns, amongst others. It was administered to a community leader.

In addition, on the day of the filming of the visit, other information was collected including the length of relationship between the home visitor and the family, total number of visitors the family had had to date, etc.

[11] The number of varieties of play materials is the sum of indicators for: toys that make/play music; toys/objects meant for stacking/constructing/building; things for drawing/writing/coloring/painting; toys for moving around; toys to play pretend games; picture books; drawing books for children; and toys for learning shapes and colors. The number of varieties of play activities is the sum of indicators for: reading books/looking at picture books; telling stories to child; singing songs with child; playing with child with toys; spending time with child scribbling/drawing/coloring; spending time with child naming things/counting; and taking child outside for a walk.

4. Samples and descriptive statistics

We work with three different samples aimed at maximizing sample size across our estimations. The three samples were selected from the sample with quality of home visits data, as shown in Figure 1 (Panel B). The sample included 229 focus children (who were also part of the impact evaluation data collection) and 176 home visitors. We excluded focus children, home visitors and any videos of their visits if: (i) we did not have identifying information for the home visitor and hence could not merge that specific case to other data sources; (ii) children did not have developmental outcomes at baseline and endline of the impact evaluation; (iii) the home visitor did not serve any child for whom we had developmental outcomes; (iv) only one home visit of the home visitor was filmed. After dropping these cases, we were left with a sample of 218 children who were visited by 155 home visitors. We refer to this sample as the 'own visit sample' as these are children for whom home visits were observed and filmed.

For the second and third samples, we used the Program's administrative data to identify other children in the treatment group of the impact evaluation sample who had received visits by any of these 155 home visitors between January and December 2015[12], even though their own visits had *not* been filmed during the quality data collection. For these children, we also have developmental outcomes at baseline (ASQ-3) and endline (ASQ-3 or Bayley-III). Specifically, the second sample includes 431 children who we could match at the home visitor level and for whom we have ASQ-3 scores—this is, 218 in the 'own visit sample' plus 213 additional children with ASQ-3 scores. The third sample includes the 346 children we were able to match and who were also assessed on the Bayley-III—this is, 218 in the 'own visit sample' plus 128 additional children with Bayley-III scores. We refer to these samples as the 'ASQ sample' and 'Bayley sample', respectively.

Table 1 illustrates some descriptive statistics on our three analysis samples[13]. Characteristics of the children and their families at baseline and endline are presented in Panels A and B, respectively. Panel B also shows information on the communities where children and their families live. Children whose home visit was filmed ('own visit sample') were, on average, 31.82 months of age at endline. As one would expect, matched children in the impact evaluation sample were slightly older since these children were not necessarily receiving visits any longer at the time of the filming (children graduate from the Program at 36 months). Children in the 'ASQ sample' were 36.26 months old and children in the 'Bayley sample' were 33.86 months old on average. Overall, there are fewer girls than boys in all three samples—the proportion of girls being 47 percent, 46 percent and 44 percent, respectively. Similarly, mothers have between 6 and 7 years of education in all three samples on average, which corresponds to completed primary and one year of secondary. Regarding the quality of their home environments—this is the FCI indices for number of varieties of play materials and of play activities—we see that, on average, children had engaged in

[12] The Program's administrative data did not have such information prior to January 2015.

[13] Araujo et al. 2018 presents a comparable set of descriptives for the full impact evaluation sample from which these sub-samples were drawn.

between 4 and 5 different activities (out of 7) in the week prior to the endline survey and used between 2 and 3 types of toys (out of 8) in their play. These averages are very similar across samples. Over half of the children in the 'own visit sample' (56 percent) had at least one other home visitor since their enrolment in the Program and between one and two more home visitors, conditional on having had more than one. They had been in the Program for about 23 months, on average.

Using administrative data on Program enrollment, we estimated the total number of visits a child should have received given child's age when first participated in Cuna Mas. We compared this to the number of visits reported as actually delivered to obtain an estimate of the intensity of the intervention—namely, the percentage of visits received from the maximum possible. Although the administrative data might be imprecise or incomplete, a close examination of the distribution of these numbers across the three samples suggest that they are largely concentrated above 80 percent, meaning that most children received an intervention of an intensity that was close to what was intended. Averages across the three samples are very similar in magnitude, at around 88-89 percent.

The communities where these visits take place are rural, small and isolated. All of them are more than one hour away from the district capital, either by car or on foot. In more than half of them walking is the most common means of transportation to get there. Access to public and private services varies greatly: 87 percent of the communities in the sample have a health center, 96 percent a public preschool, but only 11 percent have a bank. In turn, coverage of flagship social programs—including school feeding—is large in these communities.

Table 2 reports home visitor characteristics obtained from the home visitor survey and the supervisor-administered checklist. These data refer to the 155 home visitors that served all children in each of the three samples and who lived in 92 communities.[14] The home visitors are, on average, 31.44 years old and have 10.05 years of education. They work, on average, with 9.85 families and conduct 4.33 visits per family per month. While home visitors have been in the Program for close to 17 months, on average, they report having worked with the family whose visit was assessed for slightly over 14 months. This is in part because the home visitors include any training time in their total employment time, and it also signals the turnover rate amongst home visitors. Almost all home visitors in the sample (98 percent) report this as their main occupation, which demands 11.27 hours from them every week.

Twenty-six percent of home visitors report having some prior experience working with children in the 0-3 age range, and 27 percent report having worked with families before their current role. Among those who have had these relevant prior experiences, the average length of past experiences is around 2.5 years. Outside the Program, 59 percent of home visitors in the sample report participating in at least one training in the past 3 years. The most common topics of the trainings they have received are

[14]Information from the home visitor survey is only available for 136 of the 155 home visitors since this survey was not administered to all home visitors in each community due to budgetary constraints.

nutrition, hygiene and health. Themes more directly related to the spirit of the Cuna Mas intervention, such as child development, child care and early learning, are far less frequent: at most one of every five home visitors report having received some training on these subjects.

As part of their time in Cuna Mas, 90 percent of home visitors have participated in pre-service training. Additionally, they have received 3.44 in-service training sessions on average while in the Program. Ninety-three percent of home visitors recognize their interaction with their supervisors as part of their in-service training experiences. However, the frequency of this interaction is not constant. Most home visitors, 63 percent, meet with their supervisor once every 2 weeks, as mandated by Program guidelines. Nevertheless, 29 percent only do it once a month and 7 percent do it even less frequently than that (every 2 months or more).[15] Another area where we see heterogeneity in work conditions is the frequency of pay: while 74 percent of home visitors report receiving their pay monthly, 17 percent report receiving it every two months, and the remaining 9 percent even more sporadically. Given community leaders play a role in facilitating payments, these differences are not surprising.

Table 3 presents summary statistics on child development, as assessed by the ASQ-3 and Bayley-III, and on the quality of the home visits, as measured with the HOVRS-A+2. Baseline and endline ASQ-3 raw scores are computed over a total of 90 points and are similar across the three samples. Bayley-III raw scores are also of similar magnitude across the two samples for which they are available, although slightly larger for the 'Bayley sample', since children are older on average and the raw scores are increasing in age by construction.

In the last three rows of Table 3, we also report quality levels of the visits—this is, the HOVRS-A+2 total score and scores for the *Home Visitor Practices* and *Family Engagement* domains. For children in the 'own visit sample', the HOVRS-A+2 means reported correspond to the quality of their *own home visit*—this is to say, the visit that was filmed and coded (own score). For children in the 'ASQ sample' and 'Bayley sample', we report the *average value of home visit quality for all visits carried out by their home visitors* and that were filmed and coded. However, for the 218 children whose visit was filmed, the *average value of home visit quality for all visits carried out by their home visitors* excludes their own score (following the logic explained in the next section). Values are very similar across samples and reflect visits of 'adequate' quality range for both domains and for the HOVRS A+2 total score. Scores on the first domain, *Home visitor practices*, are about one point lower on average than scores on the second domain, *Family Engagement*. Rubio-Codina et al. (2018) provide an in-depth discussion of these scores.

5. Empirical methodology and results

[15] We do not have data on the training received by supervisors in the sample. As per Program guidelines, we know that supervisors are meant to receive a pre-service training of 4-5 days, and at least two training workshop per year (3-5 days each) as well as monthly refresher sessions (Josephson et al. 2017).

This section explores the association between home visiting quality, maternal education, quality of the home environment, and child development outcomes. Understanding how the home visit is implemented should shed light on what are the critical aspects for quality of the service provided, those that ought to be strengthened to improve impacts on beneficiaries.

5.1. Correlations between home visiting quality, quality of the home environment, and child outcomes

We construct a mean across visits and use this variable as a proxy of the average quality that each home visitor offers to the children she serves. In constructing this mean, we might want to exclude the actual child whose visit was observed for reasons we discuss below. Specifically, if home visitor *j* serves children *i=1,…n*, we exclude child *n* from the measure of average quality corresponding to that child. In other words, home visiting quality for child *n* is defined as:

$$Q_n = \frac{Q_1 + Q_2 + \ldots + Q_{n-1}}{n-1}$$

Two reasons explain our choice of this variable as a measure of home visiting quality for child *n.* First, if we were to use only one score of home visiting quality for each individual home visitor, it is likely that this measure would be noisy (non-representative) and therefore not capture any association between home visiting quality and child development. An alternative would be to construct an average score across all the visits that we observe for that individual home visitor and use this average score in the regressions of all the children she serves. However, this could result in omitted variables bias. It is plausible that unobserved child- or family-level variables—such as child's temperament, or mother's ability to recognize and respond to the child's needs—could be associated both to the child's level of development *and* to the easiness of the visit and ultimately its quality. Home visitor's unobserved abilities might be a better fit with children or families of specific unobserved characteristics, resulting in better quality home visits. We assume that an average measure of home visiting quality that excludes the child whose development score is being regressed on quality would not have this problem.

Table 4 shows the partial correlations between the different child development outcomes—this is ASQ-3 and Bayley-III z-scores at endline—and (i) the measures of quality of the home environment (FCI varieties of activities and varieties of play materials, the HOME, and the KIDI—all expressed in z-scores) and (ii) the domains and total score of the HOVRS-A+2 for the three samples of consideration. We present two different measures of home visit quality. For children in the 'own visit sample', we correlate, in the first three rows of the last panel of Table 4, child *i*'s developmental score with the score of the visit observed and coded for child *i* and this child's home visitor (own score); and in the bottom three rows of the panel, child *i*'s developmental score with the average score of the visits observed and coded for *other* children served by the same home visitor as *i* (average of others' visits score).

With respect to the quality of the home environment, we see that the varieties of play materials are positively and significantly correlated to higher child developmental scores for all samples and all measures except for ASQ-3 scores in the 'own visit sample'. The correlations for this measure tend to be larger with Bayley-III scores than with ASQ-3 scores. The HOME is correlated to all ASQ aggregates in the 'ASQ sample' and the 'Bayley sample', but only to the ASQ aggregate which comprises all five domains in the 'own visit sample.' It is only correlated to the Bayley-III in the 'own visit sample.' The KIDI is not correlated to any of the child development outcomes, except for one instance with the Bayley cognition and language skills aggregate in the 'own visit sample.'

The HOVRS-A+2 domains and total scores are positively and significantly correlated to endline ASQ-3 scores for all samples, but not to the Bayley-III scores. When looking separately at the two HOVRS-A+2 domains, we find that correlations between child development and *Family Engagement* are higher than those with the *Home Visitor Practices* domain.

5.2. Regression analysis: association of maternal education, home environment and home visit quality with child development scores

We explore the association between child development scores and maternal education, the home environment and home visit quality using regression analysis. For each one of the three samples, we regress the child development ASQ-3 and Bayley-III z-scores, as available, controlling for the following variables:

- In a first specification (model 1) we include the same controls as those used in the impact evaluation of Cuna Mas, reported in Araujo et al. 2018[16]. We also control for our main variable of interest: home visit quality, as assessed by the HOVRS-A+2 scores, expressed in z-score.
- In the second specification (model 2), we add the measures of the quality of the home environment—this is the endline FCI play activities and play materials varieties indices, the HOME, and the KIDI (all expressed in z-scores). We include these variables to test for an indirect effect of the HOVRS through the quality of parenting practices.

Standard errors (SEs) are clustered at the level of the district since it was the unit of randomization for the impact evaluation. We also include controls for the units of stratification used for the sample design.

Results, reported in Tables 5, 6 and 7 (one for each sample), differ across child development measures. Baseline child development outcomes are not significantly associated with endline child development on any of the models. Surprisingly, neither is maternal education. Similarly, the varieties of play activities are not associated with

[16] These correspond to the baseline child and family characteristics presented in Table 1, as well as the baseline ASQ total score (all domains) and its quartic. They also include village and district controls, namely the average at the village level of all baseline child and household controls from Table 1; the average score at the district level in reading comprehension and mathematic of the national standardized test "*Evaluación censal de Estudiantes*", and the average at the district level of household controls from the Peruvian 2007 national census.

developmental outcomes, while the varieties play materials are associated to both the ASQ-3 and the Bayley-III aggregates in the ASQ and Bayley samples, but not in the own visit sample, which has a smaller sample size. The number of visits received by the child is not statistically significant in any of the specifications.

In all regressions that use the ASQ-3 as a dependent variable, regardless of the sample and the aggregate, the HOVRS-A+2 total score is positively and significantly associated with child development, suggesting a correlation between home visit quality and child development. The magnitude of this association is not small: a 1 SD improvement in the quality of the home visit is associated to an increase of ASQ-3 scores of between 0.149 and 0.257 SD. Nonetheless, the coefficient on the HOVRS-A+2 total score is not significant at conventional levels for those regression that use the Bayley-III z-score as an outcome measure, which is inconsistent with the results on the ASQ-3.

We conducted many robustness checks, all of which lead to qualitatively comparable results. More specifically:

- *Clustering of the SEs at the home visitor level* instead of the district level does not modify results. Similarly, findings are minimally changed if we do not cluster the SEs at all.
- Estimating all models *using the HOVRS-A+2 scores at the domain and item levels* returns consistent findings throughout. Impacts are only significant when the ASQ-3 score is the dependent variable, and more items are significantly associated with child development in the *Family Engagement* domain than in the *Home Visitor Practices* domain, with regression coefficient also being larger.
- *Removing items in the HOVRS-A+2 that are not aligned with the objectives and structure of the Cuna Mas visits* does not modify our main results either.
- *Trying alternative aggregations of the quality of home visitor HOVRS-A+2 variable* does not modify findings either. Given the structure of our sample, we construct an average measure of home visitor quality based on a different number of observed visits for different home visitors. It is possible to estimate the 'own-visit model' fixing that number at the minimum common to all observations in that sample. For children in the 'own visit sample', we ran 500 draws of our main results in Table 4, and for each one we computed the average quality score for that child with a fixed number of two visits of the child's home visitor (other than the child's own score). Each draw selected at random a different pair of visits to compute the average quality score for the child. Results show that the coefficient on the HOVRS-A+2 and the standard error of the coefficient are stable across these iterations, both for the ASQ-3 and for the Bayley-III as the outcome measures.
- *Dropping out of the analysis observations where videos did not meet the HOVRS-A+2 reliability criteria after double coding* did not affect our main results.

6. Conclusions

In this paper, we have investigated the association between home visit quality and

child development using a sample of children and home visits that are part of the Cuna Mas impact evaluation. Home visit quality was assessed by the HOVRS-A+2, commonly used to evaluate quality of home visiting services in the US. The instrument was suitably adapted to the areas where Cuna Mas operates—namely, rural disperse communities in Peru. The HOVRS-A+2 was scored by trained coders on video recordings of 554 home visits, in an identical number of homes. Child development was assessed using the ASQ-3 and the Bayley-III. The ASQ-3 was administered at baseline and endline to all children in the evaluation sample, and the Bayley-III was administered to a sub-sample of these children, only at endline. For each test, an aggregated developmental z-score was constructed with the cognitive, language and fine motor scales.

In the analysis, we work with three distinct samples of children, depending on the availability of developmental outcomes and on whether we have video recordings of their own visits or of the visits carried out by their home visitors to other children. Overall, results are consistent across samples. For the three samples, filmed and scored videos correspond to home visits carried out by a total of 155 home visitors in 92 rural communities. As mandated by Cuna Mas' guidelines, the home visitors in the sample worked with between nine and ten children. They had been in the Program for almost two years, a majority had some secondary education—some completed, others not—and about a quarter had worked with families in previous employment. The children in the sample were between two and two and a half years old on average and a bit more than half of them had had at least one other home visitor since their enrolment in the Program (sometimes two). Children in the sample appear to have received above 80 percent of intended Program visits, suggesting high compliance or Program intensity.

Whilst we find that home visiting quality scores are significantly associated with the ASQ-3 z-score, controlling for child characteristics (including baseline scores), maternal education and the quality of the home environment and other controls reported in the Cuna Mas impact evaluation, the associations were not statistically significant if the Bayley-III z-score was used as an outcome. This is both inconsistent and surprising since the Bayley-III and other multi-dimensional diagnostic tests are considered to be the gold standard test for the assessment of child development (Frongillo et al. 2014; Fernald et al. 2009; Fernandes et al. 2014) and the ASQ-3 has been shown to have poorer validity when compared to other screener tests and to the Bayley-III (Rubio-Codina et al. 2016a). Also surprising are the lack of conditional associations between maternal education and the quality of the home environment—varieties of play activities, in particular—and child development scores, since these are well documented in the literature (Rubio-Codina et al. 2016b). One reason might be the small sample sizes. A second possibility is that at very low levels of education, like the ones in our sample, years of maternal education might reflect a lack of access as opposed to the mother's motivation to learn, as it often does in better-off samples. A third explanation might be that, given that our identification strategy does not allow any inference of causality, the home visit quality variable might de facto be capturing

many other child and family characteristics that are associated with (i.e., contribute to) home visit quality and better child development.

Nonetheless, the positive associations between the ASQ-3 and home visiting quality found in this paper are consistent with the theory in the literature on home visiting service delivery arguing that process quality is key to achieving the desired outcomes in participant children (Paulsell et al. 2010; Wasik and Bryant 2001). In practice, however, measures to assess home visit quality such as the HOVRS-A+2 have only recently been developed and only a handful of papers have reported associations between home visiting process quality and child development, which limits the comparisons of our results to other studies. Two studies that we are aware of include Roggman et al. (2016) and Vogel et al. (2015), both in the context of the Early Head Start program.

This indication that process quality is related to child developmental outcomes has a clear policy implication regarding the focus of the training of the home visiting workforce. Home visitors are required to master a combination of very subtle abilities, such as explicit instruction, sensitive and warm interactions, and responsive feedback to create the desired behavioural changes in beneficiaries. Yet they are often ill-educated, poorly trained for such responsibilities, and receive little or no feedback about their performance. If process quality is truly at the 'heart' of service delivery (Wasik and Bryant, 2001), further research is needed to understand which cost-effective training strategies and mentoring schemes can better support this workforce in their daily responsibilities. This is particularly important for home visiting programs such as Cuna Mas which operate at scale, in low-resource, low-skill contexts and with very disperse populations.

References

Ángeles, G., GaDEen, P., Galiani, S., Gertler, P., Herrera, A., Kariger, P. y Seira, E. (2011). "Evaluacion de Impacto del Programa Estancia Infantiles para Apoyar a Madres Trabajadoras. Informe Final de la Evaluación de Impacto." Instituto Nacional de Salud Pública. México.

Attanasio O., Fernández C, Fitzsimons E, Grantham-McGregor S, Meghir C, Rubio-Codina M. (2014) "Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial". *British Medical Journal*, 349:g5785

Araujo, M.C., Dormal, M. Grantham-McGregor, S., Lazarte, F., Rubio-Codina, M., and Schady, N. (2018). "Behavioral change and child development at scale". Unpublished manuscript.

Bayley N. (2006). Bayley Scales of Infant and Toddler Development—Third Edition: Technical manual. San Antonio, TX: Harcourt Assessment.

Berlinski, S., & Schady, N. (2015). The Early Years: Child Well-being and the Role of Public Policy. Development in the Americas Series.

Bernal, R. (2015). "The Impact of a Vocational Education Program for Childcare Providers on Children's Well-being". *Economics of Education Review*, vol. 48(C), pages 165-183.

Boller, K., Vogel, C., Cohen, R., Aikens, N., and Hallgren, K. (2009). "Home Visit Characteristics and Content Form." Princeton, NJ: Mathematica Policy Research.

Britto, Pia R, Lye. S.L., Proulx K., Yousafzai, A., Matthews, S. Vaivada, T., Perez-Escamilla, R., Rao, N., Ip, P., Fernald, L., MacMillan, H., Hanson, M., Wachs, T., Yao, H., Yoshikawa, H., Cerezo, A., Leckman, J., Bhutta, Z. and the Early Childhood Development Interventions Review Group, for the Lancet Early Childhood Development Series Steering Committee. (2016). Nurturing care: promoting early childhood development, *Lancet* http://dx.doi.org/10.1016/S0140-6736(16)31390-3

Caldwell, B., y Bradley, R. (2003). Home Observation for Measurement of the Environment: Administration manual. Little Rock, AR: University of Arkansas.

DGSE-MIDIS, DGPP-MEF, BID. (2015). "Documento de Línea de Base para la Evaluación de Impacto del Programa Nacional Cuna Más – Servicio de Acompañamiento a Familias", Lima, febrero de 2015, Documento interno del MIDIS.

Dishion, T. J., Knutson, N., Brauer, L., Gill, A., & Risso, J. (2010). Family Check Up: COACH ratings manual. Unpublished coding manual. Available from the Child and Family Center, 195 West 12th Avenue, Eugene, OR 97401.

Fernald, L., Kariger, P., Hidrobo, M. y Gertler, P. (2012). “Socioeconomic gradients in child development in very young children: evidence from India, Indonesia, Peru, and Senegal.” Proceedings of the National Academy of Science 109 (2): 17273-80.

Fernald, L. C., Kariger, P., Engle, P., & Raikes, A. (2009). Examining early child development in low-income countries. *Washington DC: The World Bank*.

Fernandes M, Stein A, Newton CR, Cheikh-Ismail L, Kihara M, Wulff K, et al. The INTERGROWTH-21st Project Neurodevelopment Package: A Novel Method for the Multi-Dimensional Assessment of Neurodevelopment in Pre-School Age Children. (2014) *PLOS ONE.* 9(11):e113360. doi: 10.1371/journal.pone.0113360 PMID: 25423589

Frongillo EA, Tofail F, Hamadani JD, Warren AM, Mehrin SF. Measures and indicators for assessing impact of interventions integrating nutrition, health, and early childhood development. (2014). *Annals of the New York Academy of Scences,* 1308(1):68–88.

Frongillo EA, Sywulka SM, Kariger P. (2003). UNICEF psychosocial care indicators project. Final report to UNICEF. Ithaca: Division of Nutritional Sciences, Cornell University.

Lazarte, F. (2016). *Informe final de las actividades realizadas para el recojo de información del estudio de Bayley III*. Reporte de consultoría no publicado.

MacPhee, D. (2002). Knowledge of Infant Development Inventory Survey of Child Care Experiences Manual. Mimeo.

MEF-BID, (2016). “Resultados de la evaluación de impacto del Servicio de Acompañamiento a Familias del Programa Nacional Cuna Más”. Unpublished manuscript.

Nahar B, Hossain MI, Hamadani JD, Ahmed T, Huda SN, Grantham-McGregor SM, (2012). “Effects of a community-based approach of food and psychosocial stimulation on growth and development of severely malnourished children in Bangladesh: a randomised trial”. *European Journal of Clinical Nutrition*; 66:701-9

Paulsell, D., Avellar, S., Martin, E. S., & Del Grosso, P. (2010). Home visiting evidence of effectiveness review: Executive summary (No. 5254a2ab30e146ce900220dbc4f41900). Mathematica Policy Research.

Roggman, L. A., Cook, G. A., Innocenti, M. S., Jump Norman, V. K., Boyce, L. K., Christiansen, K., and Peterson, C. (2016). “ Home Visit Quality Variations in two Early Head Start Programs in Relation to Parenting and Child Vocabulary Outcomes”. *Infant Mental Health Journal.* Vol. 00(0) 1-15.

Roggman, L. A., Cook, G. A., Innocenti, M. S., Jump Norman, V. K., Christiansen, K.,

Boyce, L. K., Aikens, N., Boller, K., Paulsell, D., & Hallgren, K. (2014). Home Visit Rating Scales—Adapted and Extended (HOVRS-A+) version 2. Adapted from Roggman, L. A., et al (2008). Home Visit Rating Scales. In L. Roggman, L. Boyce, and M. Innocenti, Developmental Parenting (pp. 209-217). Baltimore: Paul H. Brookes Publishing.

Roggman, L. A., Cook, G. A., Jump, V., Boyce, L. K. & Innocenti, M. S. (2006). Home Visit Rating Scales (HOVRS). Unpublished manuscript. Utah State University, Logan, UT.

Rubio-Codina, M. Araujo, M.C., Attanasio, O., Muñoz, P. Grantham-McGregor, S. (2016a). Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies. *PlosOne.* Published: August 22, 2016. http://dx.doi.org/10.1371/journal.pone.0160962

Rubio-Codina, M., Attanasio, O., & Grantham-McGregor, S. (2016b). Mediating pathways in the socio-economic gradient of child development: Evidence from children 6–42 months in Bogota. *International journal of behavioral development*, *40*(6), 483-491.

Rubio-Codina, M., Dormal, M. and Araujo, M.C. (2018). "Measuring quality and characterizing Cuna Mas home visits: validation of the HOVRS-A+2 in Peru and of a short checklist for use at-scale". Washington D.C.: Inter-American Development Bank.

Schodt, S., Parr, J., Araujo, M.C., and Rubio-Codina, M. (2015). "Measuring the Quality of Home-Visiting Services. A review of the literature". Inter-American Development Bank, Social Protection and Health Division Technical Note IDB-TN-881, October.

Smith, J. D., Dishion, T. J., Shaw, D. S., & Wilson, M. N. (2013). "Indirect effects of fidelity to the family check-up on changes in parenting and early childhood problem behaviors". *Journal of Consulting and Clinical Psychology*, 81(6), 962.

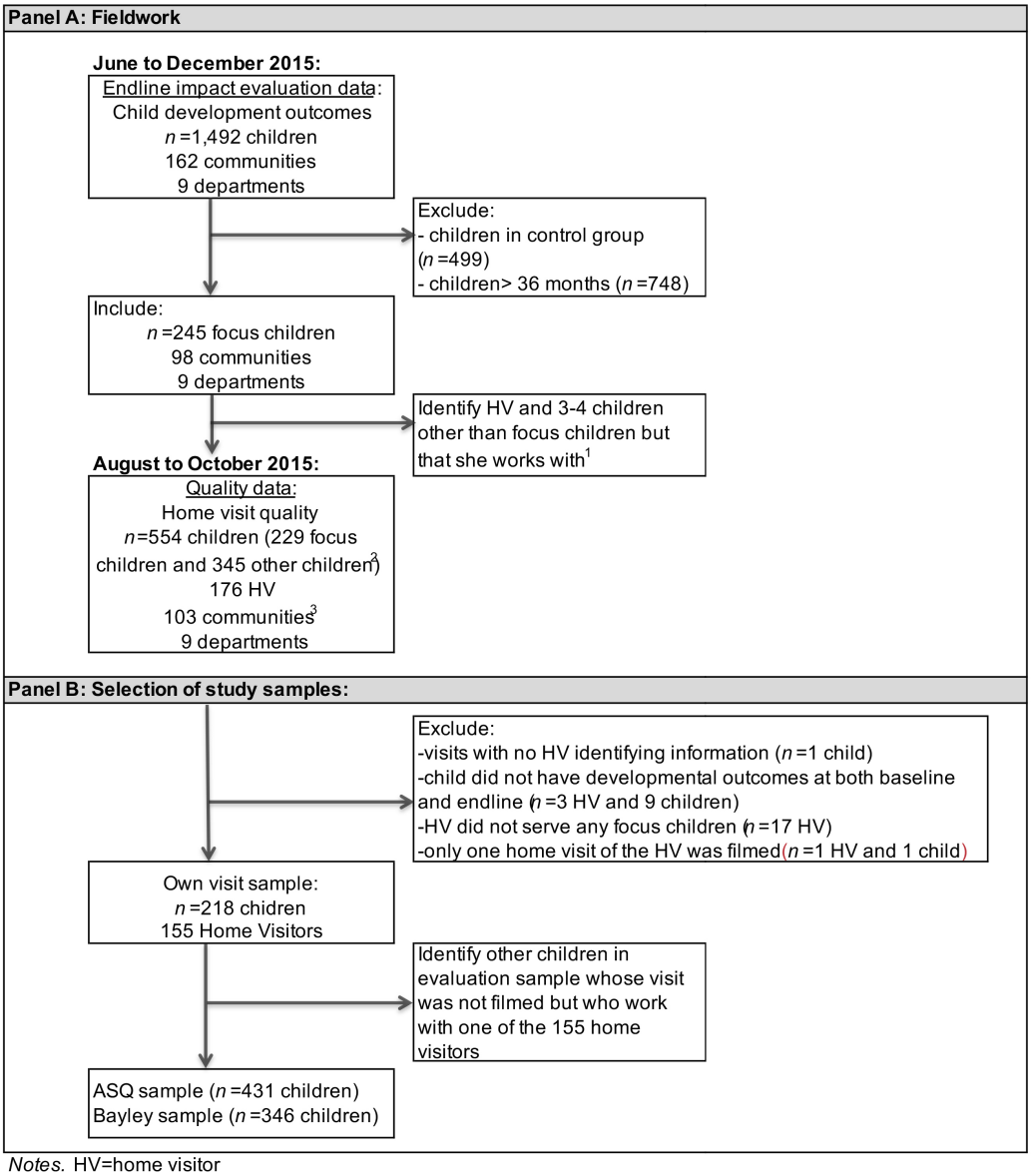
Squires J, Bricker D, Twombly E, Nickel R, Clifford J, Murphy K, et al. (2009). Ages & Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed, Child-Monitoring System. Baltimore, MD: Paul H. Brookes Publishing Co.

Vogel, C. A., Caronongan, P., Thomas, J., Bandel, E., Xue, Y., Henke, J., Aikens, N., Boller, K. & Murphy, L. (2015). Toddlers in Early Head Start: A Portrait of 2-Year-Olds, Their Families, and the Programs Serving Them. Volume I: Age 2 Report. OPRE Report 2015-10. Mathematica Policy Research. February 2015. Available at: https://www.mathematica-mpr.com/our-publications-and-findings/publications/volume-i-toddlers-in-early-head-start-a-portrait-of-2yearolds-their-families-and-the-programs

Wasik, B. H., and Bryant, D. M. (2001). *Home visiting: Procedures for helping families*, *2nd ed*. Thousand Oaks, CA: Sage.

Tables and Figures

Figure 1: Flow diagram of study participants



Notes. HV=home visitor

[1]For six home visitors (3% of the sample) we only had one filmed visit and for ten home visitors (6%) we only have two filmed visits.

[2]Includes the 229 of the 245 focus children with complete information on the HOVRS and the checklist.

[3]Includes 97 of the 98 communities that were identified, and 6 new communities not included in the evaluation sample where children other than the focus children live.

Table 1: Child, family and community characteristics, by sample

	N	Mean/ Proportion	SD	N	Mean/ Proportion	SD	N	Mean/ Proportion	SD
	Own visit sample			ASQ sample			Bayley sample		
Panel A: baseline information									
Child is Female	218	0.47		431	0.46		346	0.44	
Mother's age	218	27.50	7.71	431	28.49	7.68	346	27.97	7.57
Maternal education (years)	207	6.83	3.77	405	6.65	3.74	327	6.61	3.71
Mother is indigenous	218	0.01		431	0.02		346	0.02	
Household with both parents	218	0.89		431	0.86		346	0.88	
Household size	218	5.34	2.04	431	5.46	2.11	346	5.47	2.04
Number of bedrooms in home	218	2.42	1.26	431	2.52	1.31	346	2.48	1.31
Number of assets (0-11)	218	2.86	2.23	431	2.71	2.15	346	2.68	2.17
Household has piped water	218	0.64		431	0.67		346	0.65	
Household is connected to sewerage system	218	0.28		431	0.30		346	0.27	
Main material of floor in home is not earth	218	0.22		431	0.22		346	0.21	
Panel B: endline information									
Child and family characteristics									
Age (months)	218	31.82	2.85	431	36.26	6.37	346	33.86	4.09
Time in Program (months)[1]	217	23.18	5.99	-	-	-	-	-	-
At least one other home visitor since enrolment[1]	218	0.56		-	-	-	-	-	-
Number of home visitors since enrolment (if more than 1)[1]	120	1.84	0.96	-	-	-	-	-	-
Proportion of visits received	218	0.89		431	0.88		346	0.88	
FCI, number of varieties of play activities	218	4.47	2.09	431	4.35	2.14	346	4.36	2.16
FCI, number of varieties of play materials	215	2.32	1.71	428	2.24	1.67	343	2.21	1.63
HOME, total score (reverse coded)	218	9.99	1.82	431	9.87	1.90	346	9.89	1.87
KIDI	215	0.59	0.14	421	0.59	0.14	340	0.59	0.14
Community characteristics (N=92)[2]									
Time to the district capital on foot (hours)[3]	46	1.79	3.57						
Time to the district capital by car (hours)[3]	43	1.02	0.84						
Health center = 1	92	0.87	0.34						
Preschool center of the Ministry of Education = 1	92	0.96	0.21						
Bank or *caja municipal/rural* =1	92	0.11	0.31						
School breakfast program "Qali Warma" = 1	92	0.97	0.18						

Notes: [1]Data from the supervisor-administered checklist and as such only available for children in the quality study. All other variables come from the household or community survey.

[2] The community charateristics are the same for all three samples.

[3]Only for communities that reported it as the principal mean of transportation.

Table 2: Home visitor characteristics (N=155), all samples

	N	Mean/ Proportion	SD
Age (years)	136	31.44	8.09
Years of education	136	10.05	2.52
Time in Program (months)[1]	155	16.79	9.69
Time working with family (months)[1]	155	14.45	9.52
Number of families she visits (per month)	136	9.85	2.89
Number of visits per family (per month)	136	4.33	2.06
Proportion for whom the main occupation is home visitor (previous week)	136	0.98	
Number of hours she works as a home visitor (per week)	136	11.27	5.12
Proportion with experience with children in the 0-3 age range	136	0.26	
Years of experience with children in the 0-3 age range	35	2.38	3.08
Proportion with experience working with families	136	0.27	
Years of experience with working with families	37	2.52	3.39
Poportion that have been trained in the last 3 years	136	0.59	
Porportion that have been trained in:			
Child care	136	0.19	
Hygiene	136	0.49	
Health	136	0.41	
Nutrition	136	0.42	
Child development	136	0.21	
Early learning	136	0.14	
Proportion that were trained when they entered the Program	136	0.90	
Number of trainings received (in additional to training when entered Program)	136	3.44	2.53
Proportion that are trained by the supervisors during visits	136	0.93	
Frequency of training by supervisor			
biweekly	126	0.63	
monthly	126	0.29	
bimonthly/other	126	0.07	
Pay frequency			
monthly	136	0.74	
bimonthly	136	0.17	
trimestrial/other	136	0.09	
KIDI score (proportion of correct answers)	136	0.75	0.09

Notes: [1]Data from the supervisor-administered checklist. Other variables com from the home visitor survey.

Table 3: ASQ, Bayley and HOVRS-A+2 raw scores, by sample

	N	Mean/ Proportion	SD	N	Mean/ Proportion	SD	N	Mean/ Proportion	SD
	Own visit sample			ASQ sample			Bayley sample		
ASQ-III raw scores									
Baseline									
ASQ total (all domains)	218	45.29	11.48	431	45.27	11.38	346	45.60	11.18
ASQ, cognition, language, and fine motor skills	218	47.45	13.34	431	46.42	13.32	346	47.65	12.81
ASQ, cognition and language skills	218	46.67	13.86	431	46.38	14.21	346	47.70	13.49
Endline									
ASQ total (all domains)	218	43.54	9.10	431	43.76	9.68	346	43.42	9.58
ASQ, cognition, language, and fine motor skills	218	40.47	8.67	431	40.49	9.69	346	40.46	9.36
ASQ, cognition and language skills	218	40.30	10.31	431	40.15	11.01	346	40.25	11.04
Bayley-III raw scores									
Bayley total (all domains)	218	42.34	3.16	-	-	-	346	43.49	3.51
Bayley, cognition, language, and fine motor skills	218	46.38	3.20	-	-	-	346	47.46	3.62
Bayley, cognition and language skills	218	48.27	3.25	-	-	-	346	49.24	3.54
HOVRS-A+2 raw scores[1]									
HOVRS, Home Visitor Practices domain	218	3.43	0.52	431	3.44	0.42	346	3.43	0.43
HOVRS, Family Engagement domain	218	4.47	0.73	431	4.50	0.56	346	4.49	0.56
HOVRS, all domains	218	3.88	0.56	431	3.90	0.45	346	3.88	0.46

Notes: [1] In own visit sample, score of the child's visit; in ASQ and Bayley samples, mean score of all visits of the child's home visitor, excluding the child's own score.

Table 4: Pearson correlations between the ASQ and the Bayley with the quality of the home environment and home visiting quality, by sample

	Quality of the home environment/ Parental Knowledge				Home visiting quality (HOVRS)					
	FCI play activities	FCI play materials	HOME	KIDI	HVP[1]	FE[1]	Total[1]	HVP[2]	FE[2]	Total[2]
Panel A: Own visit sample (N=218)										
ASQ total (all domains)	0.05	0.16**	0.20***	0.07	0.09	0.22***	0.18***	0.11*	0.20***	0.17**
ASQ, cognition, language, and fine motor skills	-0.03	0.05	0.08	0.04	0.12*	0.20***	0.18***	0.10	0.17**	0.14**
ASQ, cognition and language skills	-0.02	0.02	0.08	-0.01	0.17**	0.24***	0.22***	0.13*	0.19***	0.17**
Bayley total (all domains)	0.10	0.17**	0.11*	0.09	0.00	0.16**	0.09	0.05	0.10	0.08
Bayley, cognition, language, and fine motor skills	0.10	0.14**	0.11	0.10	0.00	0.16**	0.09	0.06	0.11	0.09
Bayley, cognition and language skills	0.09	0.17**	0.14**	0.12*	0.00	0.16**	0.10	0.07	0.10	0.09
Panel B: ASQ sample (N=431)										
ASQ total (all domains)	0.10*	0.22***	0.20***	0.09	-	-	-	0.12**	0.21***	0.17***
ASQ, cognition, language, and fine motor skills	0.03	0.09*	0.13**	0.05	-	-	-	0.13**	0.20***	0.18***
ASQ, cognition and language skills	0.04	0.06	0.10*	0.01	-	-	-	0.16***	0.22***	0.21***
Panel C: Bayley sample (N=346)										
ASQ total (all domains)	0.10*	0.22***	0.20***	0.09	-	-	-	0.12**	0.21***	0.17***
ASQ, cognition, language, and fine motor skills	0.03	0.09*	0.13**	0.05	-	-	-	0.13**	0.20***	0.18***
ASQ, cognition and language skills	0.04	0.06	0.10*	0.01	-	-	-	0.16***	0.22***	0.21***
Bayley total (all domains)	0.10*	0.15***	0.05	0.03	-	-	-	0.03	0.07	0.05
Bayley, cognition, language, and fine motor skills	0.10*	0.14***	0.05	0.03	-	-	-	0.03	0.07	0.05
Bayley, cognition and language skills	0.06	0.14***	0.06	0.03	-	-	-	0.05	0.08	0.07

Notes: Endline scores. HVP=Home Visitor Practices; FE=Family Engagement. Coefficients significant at * p<0.10, ** p<0.05, *** p<0.01.

[1]score of the child's visit, net of coder effect

[2]mean score of all visits of the child's home visitor, excluding the child's own score and net of coder effect

Table 5: Association between child development and quality of the home visit, own visit sample (N=218)

	ASQ total (all domains)		ASQ, cognition, language, and fine motor skills		ASQ, cognition and language skills		Bayley total (all domains)		Bayley, cognition, language, and fine motor skills		Bayley, cognition and language skills	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Child is female	0.212	0.190	0.271*	0.279	0.163	0.145	0.191	0.161	0.122	0.098	0.190	0.153
	(0.152)	(0.165)	(0.161)	(0.168)	(0.141)	(0.152)	(0.152)	(0.156)	(0.148)	(0.150)	(0.159)	(0.159)
Maternal education (years)	0.027	0.020	-0.001	-0.005	-0.026	-0.026	0.004	-0.003	0.013	0.007	0.002	-0.011
	(0.027)	(0.026)	(0.038)	(0.035)	(0.035)	(0.033)	(0.039)	(0.041)	(0.039)	(0.041)	(0.040)	(0.042)
ASQ baseline total (all domains)	-0.083	-0.073	-0.085	-0.065	-0.129	-0.121	-0.122	-0.119	-0.110	-0.112	-0.161	-0.146
	(0.140)	(0.147)	(0.171)	(0.173)	(0.164)	(0.170)	(0.157)	(0.168)	(0.153)	(0.161)	(0.167)	(0.172)
Number of visits received	-0.005	-0.005	-0.002	-0.003	-0.001	-0.001	0.002	0.002	0.001	0.001	0.004	0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)
HOVRS total z-score[1]	0.198**	0.191**	0.196*	0.175	0.214*	0.206*	0.066	0.058	0.066	0.054	0.016	0.007
	(0.090)	(0.089)	(0.115)	(0.111)	(0.123)	(0.119)	(0.067)	(0.068)	(0.071)	(0.072)	(0.077)	(0.071)
FCI Number of varieties of play activities z-score		-0.015		-0.055		-0.035		-0.007		-0.006		-0.021
		(0.076)		(0.080)		(0.087)		(0.095)		(0.093)		(0.103)
FCI Number of varieties of play materials z-score		0.074		-0.052		0.024		0.054		0.020		0.108
		(0.110)		(0.118)		(0.107)		(0.085)		(0.081)		(0.093)
HOME z-score (reverse coded)		0.118		-0.005		0.025		0.070		0.078		0.114
		(0.088)		(0.085)		(0.091)		(0.089)		(0.086)		(0.090)
KIDI z-score		0.037		0.078		0.012		0.038		0.042		0.097
		(0.077)		(0.081)		(0.095)		(0.064)		(0.065)		(0.068)

Notes: Endline outcomes expressed as z-scores. Other controls include the baseline child and family characteristics presented in Table 1, the baseline ASQ total score (all domains) and its quartic, the average at the village level of all baseline child and household controls from Table 1, the average score at the district level in reading comprehension and mathematic of the national standardized test "Evaluación censal de Estudiantes", and the average at the district level of household controls from the Peruvian 2007 national census. Coefficients significant at * p<0.10, ** p<0.05, *** p<0.01.

Table 6: Association between child development and quality of the home visit, ASQ sample (N=431)

	ASQ total (all domains)		ASQ, cognition, language, and fine motor skills		ASQ, cognition and language skills	
	(1)	(2)	(1)	(2)	(1)	(2)
Child is female	0.126	0.097	0.200**	0.175*	0.153*	0.124
	(0.085)	(0.090)	(0.094)	(0.095)	(0.091)	(0.090)
Maternal education (years)	0.027	0.017	0.024	0.018	0.014	0.008
	(0.020)	(0.019)	(0.021)	(0.021)	(0.022)	(0.021)
ASQ baseline total (all domains)	0.053	0.066	-0.037	-0.019	-0.117	-0.097
	(0.106)	(0.108)	(0.112)	(0.112)	(0.111)	(0.111)
Number of visits received	-0.001	-0.001	-0.000	-0.000	-0.002	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
HOVRS total z-score[1]	0.149**	0.152***	0.157**	0.160**	0.187***	0.189**
	(0.056)	(0.056)	(0.065)	(0.065)	(0.069)	(0.071)
FCI Number of varieties of play activities z-score		0.077		-0.011		-0.000
		(0.053)		(0.052)		(0.053)
FCI Number of varieties of play materials z-score		0.148**		0.093*		0.095**
		(0.064)		(0.055)		(0.047)
HOME z-score (reverse coded)		0.077		0.043		0.067
		(0.074)		(0.060)		(0.058)
KIDI z-score		-0.007		0.009		0.009
		(0.045)		(0.041)		(0.051)

Notes: Endline outcomes expressed as z-scores. Other controls include the baseline child and family characteristics presented in Table 1, the baseline ASQ total score (all domains) and its quartic, the average at the village level of all baseline child and household controls from Table 1, the average score at the district level in reading comprehension and mathematic of the national standardized test "Evaluación censal de Estudiantes", and the average at the district level of household controls from the Peruvian 2007 national census. Coefficients significant at * p<0.10, ** p<0.05, *** p<0.01.

Table 7: Association between child development and quality of the home visit, Bayley sample (N=346)

	ASQ total (all domains)		ASQ, cognition, language, and fine motor skills		ASQ, cognition and language skills		Bayley total (all domains)		Bayley, cognition, language, and fine motor skills		Bayley, cognition and language skills	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Child is female	0.051	-0.004	0.140	0.107	0.070	0.030	0.115	0.087	0.061	0.039	0.108	0.074
	(0.095)	(0.102)	(0.107)	(0.112)	(0.107)	(0.111)	(0.116)	(0.116)	(0.115)	(0.116)	(0.119)	(0.117)
Maternal education (years)	0.020	0.010	0.018	0.013	0.010	0.005	0.012	0.010	0.015	0.014	0.010	0.006
	(0.024)	(0.023)	(0.026)	(0.025)	(0.026)	(0.025)	(0.024)	(0.025)	(0.025)	(0.026)	(0.026)	(0.027)
ASQ baseline total (all domains)	0.072	0.098	0.004	0.028	-0.053	-0.035	0.056	0.072	0.047	0.057	0.037	0.064
	(0.131)	(0.133)	(0.150)	(0.150)	(0.144)	(0.142)	(0.111)	(0.118)	(0.105)	(0.110)	(0.116)	(0.120)
Number of visits received	-0.001	-0.001	-0.000	0.000	-0.002	-0.001	0.002	0.002	0.001	0.001	0.004	0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
HOVRS total z-score[1]	0.207***	0.217***	0.224***	0.226***	0.254***	0.257***	0.036	0.041	0.041	0.046	0.003	0.005
	(0.067)	(0.072)	(0.078)	(0.079)	(0.084)	(0.085)	(0.062)	(0.065)	(0.066)	(0.069)	(0.068)	(0.069)
FCI Number of varieties of play activities z-score		0.045		-0.032		0.012		-0.022		-0.009		-0.055
		(0.056)		(0.055)		(0.055)		(0.064)		(0.066)		(0.063)
FCI Number of varieties of play materials z-score		0.200***		0.104		0.125**		0.120**		0.102*		0.140**
		(0.068)		(0.068)		(0.057)		(0.057)		(0.053)		(0.061)
HOME z-score (reverse coded)		0.110		0.045		0.042		-0.008		-0.008		0.011
		(0.070)		(0.067)		(0.072)		(0.066)		(0.063)		(0.070)
KIDI z-score		0.021		0.016		0.007		-0.017		-0.019		0.006
		(0.054)		(0.050)		(0.053)		(0.064)		(0.060)		(0.061)

Notes: Endline outcomes expressed as z-scores. Other controls include the baseline child and family characteristics presented in Table 1, the baseline ASQ total score (all domains) and its quartic, the average at the village level of all baseline child and household controls from Table 1, the average score at the district level in reading comprehension and mathematic of the national standardized test "Evaluación censal de Estudiantes", and the average at the district level of household controls from the Peruvian 2007 national census. Coefficients significant at * p<0.10, ** p<0.05, *** p<0.01.