Pre-Results Review (Registered Report):
Stage 1 Submission

Identifying Scalable and Cost-Effective Approaches to Improving Parenting Practices in Developing Settings: Experimental Evidence from India*

Irma Arteaga† Andreas de Barros‡ Alejandro J. Ganimian§

May 19, 2023

Abstract

Home-visitation programs have been shown to improve child development outcomes in low- and middle-income countries. Yet, such programs are still costly and difficult to scale due to their reliance on trained workers. This registered report presents a randomized trial of an inexpensive, low-tech, and easily scalable intervention that leverages audio recordings delivered by phone to provide guidance to mothers on how to offer psychosocial stimulation to children aged 6 to 30 months. The sample includes 2,433 mothers from 250 public childcare centers in Uttarakhand, India. Random assignment produced randomly equivalent groups as expected, and implementation fidelity and take-up up until the final weeks of the program remain high. The study will contribute to understanding whether technology can be used to expand state capacity in developing settings in an area crucial to human capital formation.

Keywords: early childhood development; education technology; state capacity; India.

JEL codes: C93; I21; I25; I38; J13.

Timeline: The expected date for completion is August 31, 2023.

*This document follows the "reporting checklist" of the Journal of Development Economics (JDE) pre-results review process (Stage 1). The study will be pre-registered at the AEA RCT Trial Registry. It was approved by the Internal Review Boards at New York University and the Institute for Financial Management and Research. We gratefully acknowledge funding from The Agency Fund and the Institute for Human Development and Social Change at New York University. We thank Sindhuja Jeyabal and Sneha Sheth for making this study possible. We also thank Arja Dayal and Rashi Maheshwari, who provided excellent research assistance. Aronee Ghosh provided helpful input throughout this process. All views expressed are those of the authors and not of any of the institutions with which they are affiliated. The authors have no conflicting interests.

†Associate Professor, Truman School of Public Affairs, University of Missouri. E-mail: arteagai@missouri.edu.

‡Postdoctoral Associate, Department of Economics, Massachusetts Institute of Technology. E-mail: debarros@mit.edu.

§Assistant Professor of Applied Psychology and Economics, Steinhardt School of Culture, Education, and Human Development, New York University. E-mail: alejandro.ganimian@nyu.edu.

1 Introduction

1.1 Background and relevance of the study

Children's early interactions with their caregivers have lasting impacts on their life outcomes. During the first years of life, vital development occurs in multiple domains (CISCD, 2000). Specifically, brain development in some domains (e.g., seeing and hearing) starts earlier and sets the foundation for others (e.g., receptive language and cognitive functions), and disruptions can impact the brain's structure and function (Grantham-McGregor et al., 2007). Environmental factors, including maternal caregiving, can catalyze or delay this process, affecting children's cognitive and emotional development (Young, 2002; Landry et al., 2006), and in turn their schooling and productivity as adults (Psacharopoulos and Patrinos, 2004).

Children in low- and middle-income countries (LMICs) are in particular need of interactions with their caregivers that are developmentally appropriate (i.e., match their emerging skills). They are disproportionately likely to face risk factors that may disrupt their development. According to a recent estimate (Lu et al., 2016), 249 million young children in these settings are exposed to two widely measured such factors—growth stunting and poverty—and are therefore at risk of not reaching their developmental potential (see also Walker et al., 2007). Interactions between caregivers and children can ameliorate the deleterious effects of these factors by promoting neurocognitive processing and brain functioning (Engle et al., 2007).

Interventions that encourage mothers to provide psychosocial stimulation to their children have improved development, school performance, and labor-market outcomes in LMICs. Most famously, a program in Kingston, Jamaica in which community health aides visited the mothers of 129 stunted children ages 9-24 months to facilitate weekly play sessions at home impacted development outcomes after two years (Grantham-McGregor et al., 1991). By age 17-18, those who had been randomly assigned to the intervention performed better on fluid intelligence and language development than their control peers (Walker et al., 2005). And 20 years after the program, the wages of its beneficiaries were 25% higher than those of the control group and on par with those of a non-stunted group (Gertler et al., 2014).

Replicating the success of this intervention at scale, however, has proven to be challenging. In recent years, many have sought to promote early stimulation through various modalities, including conditional cash transfers in Colombia (Attanasio et al., 2014), a health program in Sindh, Pakistan (Yousafzai et al., 2014), a program for pregnant and vulnerable women in Colombia (Attanasio et al., 2018), home visits in urban areas of Odisha, India (Andrew et al., 2020), and mother group sessions in rural Odisha (Grantham-McGregor et al., 2020). These delivery mechanisms have boosted children's cognitive, language, and motor skills, but they require staff to engage with mothers, which is both time-consuming and costly.

In this study, we experimentally evaluate an inexpensive, low-tech, and easily scalable approach to fostering early psychosocial stimulation: audio recordings delivered by phone. We partner with an Indian nonprofit (*Dost*) to offer this intervention to mothers of children ages 6-30 months in the state of Uttarakhand who benefit from the public childcare system (the Integrated Child Development Services or ICDS)—the world's largest such program.[1] These mothers are already supposed to receive regular home visits from the workers in their local childcare centers (known as *anganwadis* or "courtyard shelters"). Yet, as it has been documented elsewhere (Ganimian et al., 2023), *anganwadi* workers are severely overburdened. They are expected to complete more than 21 tasks across health, nutrition, and education. The audio recordings, which leverage global evidence on early psychosocial stimulation,[2] offer the government an opportunity to expand its capacity to improve child development, while ensuring that mothers receive consistent support to interact with their young children. Evidence from other LMICs suggests this approach holds promise (Arteaga and Trias, 2021). If successful, this approach could be expanded to the rest of the state and country.

Our study adds to evidence on improving early childhood development in LMICs at scale. Expansions in access to public preschool have improved learning outcomes in upper-middle income countries (Berlinski et al., 2008, 2009), but not in lower-income settings, where the capacity for public-service delivery is constrained (Bouguen et al., 2014; Blimpo et al., 2022). In these contexts, private providers have had more success (Dean and Jayachandran, 2019; Martinez et al., 2017) but only cater to a relatively small segment of the population. Our study would build on similar efforts (e.g., Ganimian et al., 2023) to expand the reach of the public sector to improve early learning outcomes in settings with limited resources.

1.2 Research questions

The study seeks to answer the following research questions:

1. Do mothers of young children sustain their interest in receiving audio recordings delivered by phone offering guidance on how to provide psychosocial stimulation?
2. Do these recordings change mothers' beliefs?
3. Do these recordings change mothers' interactions with their children?
4. Do these recordings change children's overall and language development?
5. Do these recordings change mothers' anxiety and self-efficacy?

[1] Nationally, ICDS serves over 46 million children ages 0 to 3 and another 36 million children ages 3 to 6.
[2] We describe the intervention in greater detail in section 2.4.

6. Are these recordings more effective among female children and children with low baseline levels of development?

To our knowledge, ours is one of the first randomized evaluations of audio recordings delivered by phone to promote psychosocial stimulation for young children in an LMIC. Thus, our main objective is to understand whether there is demand for such an intervention and whether it succeeds in shifting mothers' beliefs and interactions with their children.[3]

Encouraging results on mother's anxiety would validate the intervention's theory of change (see section 2.5), and positive effects on children's overall and language development would demonstrate that the elicited changes in parenting are meaningful enough to warrant further randomized expansions of the intervention to measure child-level outcomes more precisely.

We hypothesize that the intervention works primarily by raising mothers' knowledge of productive child-rearing practices (see sections 2.2 and 2.5), so evidence of heterogeneity would allow us to decide for whom to extend the intervention to maximize its impact.

2 Research design

2.1 Basic methodological framework

This is a randomized evaluation with a waitlist design. We randomly assigned 2,433 mothers in 250 *angangwadi* centers who expressed interest in receiving audio recordings by phone to either receive the intervention immediately ("treatment group") or receive it after the study ("control group"). We randomly assigned mothers within each center. Each mother had an equal probability of being assigned to the treatment or control groups within her center.

Random assignment of mothers to the intervention allows us to study the causal effect of being assigned to the intervention (the intent-to-treat, or "ITT" effect) and the effect of being assigned to the program and participating in all its phone calls (the Average Causal Response, or "ACR", scaled to the full dosage level).

2.2 Hypotheses

Our hypotheses regarding each of the research questions in section 1.2 are as follows:

1. Mothers will accept the intervention calls, they will listen to their full content, and they will answer the questions at the end of each call correctly.

[3] We expand on this discussion in section 4.

2. Mothers will change their beliefs about what constitutes developmentally appropriate interactions with their children.

3. Mothers will change the ways in which they interact with their children, in accordance with intervention guidance and their shifts in beliefs.

4. Children will improve their overall and language development.

5. Mothers will be less anxious and more self-efficacious.

6. The intervention will be more effective for mothers of female children and children with low baseline levels of development.

We discuss each of these hypotheses in greater detail, as well as how we see them come together in section 2.5, which presents our theory of change for the intervention. We present our priority order of hypothesis tests in section 3.3.

2.3 Measurement

We measure mother and child outcomes relying on mothers' self-reports by phone to keep data-collection costs manageable while we understand whether there is sustained demand for this intervention, whether it can be implemented with fidelity, whether it shifts mothers' beliefs and behaviors, and whether if affects child development. If this study demonstrates that these conditions are satisfied, we plan to use more involved and costly child assessments to measure changes in child development more precisely in subsequent studies.

We leverage existing evidence on how to administer measures of mother and child outcomes reliably and validly over the phone, which has grown rapidly in recent years in light of the Covid-19 pandemic (see Kopper and Sautmann, 2020). We use instruments that were previously administered by phone in LMICs (India, if possible) and achieved a high degree of internal-consistency reliability (Cronbach's alphas above 0.8).

RQ1: Intervention take-up. We describe program implementation and take-up by measuring the number of phone calls made to each mother, the number of calls that she accepted, the time that each mother spent on each call (in minutes), and the percentage of questions at the end of calls that the mother answered correctly throughout the 18 weeks of the intervention (for a description of these questions, see section 2.4).[4]

[4]Not all questions that mothers are asked to answer can be categorized as correct or incorrect. Some ask mothers to indicate whether they learned something new. For this purpose, we focus only on those questions assessing mothers' understanding of the material in the recordings.

RQ2: Mothers' beliefs. We measure mothers' beliefs about child development using an adaptation of the Knowledge of Infant Development Inventory (KIDI-SF; MacPhee, 1981) at endline. The KIDI-SF contains 20 statements and asks each mother whether she agrees with a statement about child development. For example, one question is "If you punish children for doing something naughty, it is okay to give them a piece of candy to stop the crying." Mothers can respond by indicating agreement, disagreement, or stating that they are not sure. This instrument has already been administered in India (Karuppannan et al., 2020). Given that it was originally developed in the 1970s, we included new statements related to screen time in children ages two and younger following recommendations from the American Academy of Pediatrics. We estimate each mother's score using item response theory and a generalized partial credit model. We standardize scores with respect to the control group at endline.

RQ3: Mother-child interactions. We measure mothers' interactions with their children with the play sub-scale of the Family Care Indicator (FCI; Hamadani et al., 2010) at baseline and endline. The FCI-play contains six items and asks each mother whether she or her child has engaged in specific activities during the week prior to the survey. For example, one question is "Have you told stories to the child last week?" Mothers can respond affirmatively or negatively. The FCI has been administered in many LMICs, including India (Grantham-McGregor et al., 2020; Luoto et al., 2021). The play sub-scale includes six items that have been previously administered on their own (Arteaga and Trias, 2021; Tofail et al., 2013; Babikako et al., 2022; Knauer et al., 2016). We estimate each mother's score using item response theory and a two-parameter logistic model. We standardize scores with respect to the control group at baseline.

RQ4: Children's overall development. We measure overall child development with the Caregiver Reported Early Childhood Development (CREDI; McCoy et al., 2017) at baseline and endline. The CREDI asks each mother whether her child can do something that they ought to be able to do, given their age. For example, for children ages 6 to 11 months, one question is "can the child pick up a small object (e.g., a small toy or stone) using just one hand?" Mothers can respond affirmatively, negatively, or by stating that they do not know. We use the short form of the CREDI, which produces a single score of overall child development. This form was validated with more than 8,000 children across 17 LMICs including India (McCoy et al., 2018; Waldman et al., 2021). Specifically, the short form contains 20 items that vary by six-month age brackets (i.e., 6-11 months, 12-17 months, etc.).

We estimate each child's score following the scoring manual (McCoy et al., 2018; Waldman et al., 2021).[5] We standardize scores with respect to the control group at baseline.

RQ4: Children's language development. We measure children's language development with an adapted version of 50 words/sentences of the MacArthur-Bates Communicative Development Inventory (CDI; Fenson et al., 2000; Jackson-Maldonado et al., 2013) at baseline and endline. The CDI asks each mother whether her child can understand and/or state a word or sentence. For example, for children ages 8 to 17 months, one question is "can the child understand and/or say *uh-oh*?" Mothers can respond indicating whether the child can understand the prompt, understand and state it, or cannot do either. Following other adaptations (e.g., Kern, 2007; Floccia et al., 2018), we translated the short form of the English CDI to Hindi and consulted with native speakers to ensure that the list of words presented to mothers is culturally relevant. We adapted the form for children in three age groups: 6-17, 18-30, and 31-37 months. We estimate each mother's score using item-response theory and a generalized partial credit model. We standardize scores with respect to the control group at baseline.

RQ5: Mothers' anxiety. We measure mothers' anxiety using the General Anxiety Disorder (GAD-7; Löwe et al., 2008) at baseline and endline. The GAD-7 asks each mother whether she has been bothered by a set of feelings during the two weeks prior to the survey. For example, one question asks "how often have you felt nervous, anxious, or on edge?" Mothers can respond using a four-point scale, from 1 ("not at all") to 4 ("nearly every day"). The GAD-7 contains seven questions and has already been administered in India (De Man et al., 2021). We estimate each mother's score using item-response theory and a generalized partial credit model. We standardize scores with respect to the control group at baseline.

RQ5: Mothers' self-efficacy. We measure mothers' self-efficacy using selected items from the Tools of Parents Self-Efficacy (TOPSE; Kendall and Bloomfield, 2005) at endline. The TOPSE asks each mother whether she agrees with a statement about her perceived capacity to engage in a parenting behavior. For example, one item is "I can recognize when my child is happy or sad." Mothers can respond choosing a number from 0 (indicating that they completely disagree with the statement) to 10 (indicating that they completely agree with the statement). Following List et al. (2021), we use items for six of the eight sub-scales, which seem to be more relevant. Using results from a validation study in Bangladesh (Ferdowshi et al., 2021), we use the two items with the highest item-total correlation from each sub-scale

[5]We use the manual's accompanying *R* package to generate the raw scaled (factor) score that reflects a child's overall development.

to construct a short version of the instrument.[6] We estimate each mother's score using item response theory and a generalized partial credit model. We standardize scores with respect to the control group at endline.

RQ6: Heterogeneous effects. We collected data on children's sex at baseline, so we can investigate whether the program has positive effects among girls.

Other variables. We collected additional data at baseline to describe our sample, including children's age (in months) and household assets. We construct a proxy measure of household poverty using the items on assets to generate an inverse covariance matrix-weighted index. We also record the number of call attempts needed to administer the survey of mothers at baseline to characterize their responsiveness to phone calls within our sample.

2.4 Intervention

2.4.1 Details of the intervention

The intervention consists of 85 audio recordings delivered by phone to mothers of children ages 6-30 months who benefit from the Integrated Child Development Services or ICDS.[7] The recordings will be delivered over 18 weeks, averaging nearly five messages per week.

The content is based on global and domestic evidence on how to improve child development. It draws on multiple international frameworks, including the Reach Up and Learn home visits in Jamaica (Chang-Lopez et al., 2020), the United Nations Children's Education Fund program guidance for early childhood development (UNICEF, 2017), the Center on the Developing Child at Harvard University's theory of change for building adult capabilities to improve child outcomes (CDC, 2011), and a text-message program evaluated by the Stanford Center for Education Policy Analysis (Cortes et al., 2021; Doss et al., 2019; York et al., 2019). It covers all themes in the National Council of Educational Research and Training's resource handbook for early childhood care and education (NCERT, 2019) and it was adjusted based on over 1,000 interviews with Indian parents to ensure the content aligns with local needs.

The recordings aim to improve children's language, cognitive, and social-emotional skills. They are organized into 18 modules: (a) the importance of early years of development; (b)

[6] We thank Sally Kendall for encouraging us to pursue this strategy.

[7] *Dost* has chosen to deliver these recordings through the phone because other forms of communication (e.g., WhatsApp) require smartphones, which are less prevalent among low-income households in LMICs. In 2021, only 52% of households with low levels of parental education in rural India had a smartphone (ASER, 2021). Further, more prevalent forms of communication (e.g., text messages) are often ignored by recipients (Beam et al., 2022).

embedding talk, care, and play into everyday life; (c) using art as a medium for learning; (d) setting up the home for learning; (e) managing screen time; (f) enabling learning through expeditions; (g) building an emotional bond; (h) creating an emotionally secure environment; (i) caring for parental well being; (j) managing difficult behavior; (k) narrating stories and having conversations; (l) supporting abilities through growth periods; (m) fostering deep and secure sibling bonds; (n) understanding nutritional relationships; (o) learning independence, empathy, and responsibility; (p) promoting physical development through play; (q) imparting experiential learning; and (r) a review of important concepts. Each module typically takes four recordings, which are offered during the same week.

Each recording (e.g., managing conflict among siblings) follows the same four-part structure. It begins by introducing a challenge that mothers may be facing and empathizes with them (e.g., siblings often fight with each other, even when mothers wish that they did not). Then, it proposes some activities for addressing the challenge at hand in everyday life (e.g., how to discipline a child who is misbehaving without comparing them to their sibling). Next, it reviews common strategies across activities (e.g., remembering each child is unique, focusing on praise, and knowing when to intervene or let siblings work things out). Lastly, it asks mothers to check their understanding or provide feedback via a touch-tone response (e.g., asking mothers to press 1 if they learned anything new about managing sibling conflict).[8]

Both descriptive and experimental evidence suggest the intervention is likely to be effective. Since its founding in 2017, the nonprofit that developed it (*Dost*, which translates to "friend") has reached 100,000 beneficiaries through government partnerships in four Indian states. A third-party survey of these beneficiaries found that 60% of those who signed up for the program became highly engaged users, 91% said that they were more confident as parents, and 94% reported having more knowledge on how to manage their children's behavior. Further, a randomized evaluation of a similar program in Guatemala found that it increased interactions between mothers and their children, decreased maternal anxiety, and improved children's vocabulary after only two months of exposure (Arteaga and Trias, 2021).

Dost recruited mothers to participate in the study with support from *anganwadi* workers. Volunteers visited *anganwadi* centers, introduced the intervention to workers, and solicited their support to enroll mothers in the study by calling an automated phone number. If a mother was randomly assigned to participate in the program, she started receiving calls. If she was not assigned, she will only start receiving intervention calls at the end of the study.

[8]These questions are asked for 75% of the calls, towards the end of each call, and they enquire about mothers' actions (e.g., "do you share your childhood stories and lullabies with your child?") and beliefs (e.g., "do you think children can learn through play?").

Additionally, *Dost* conducts "live" (i.e., non-pre-recorded) calls to keep mothers engaged with the intervention. These calls are made every Sunday to mothers who have not answered any calls in two weeks and who have not received a live call in the past month.

2.4.2 Randomization strategy

We will estimate the impact of the intervention by capitalizing on the exogenous variation in intervention assignment produced by our randomization. As discussed in section 2.1, we randomly assigned individual mothers within each *anganwadi* center to either receive the intervention immediately ("treatment group") or receive it after the study ("control group"). As we also explained in that section, each mother had an equal probability of being assigned to the treatment or control groups within her center.

Following Banerjee et al. (2020), we randomly assigned mothers to experimental groups multiple times to ensure these groups are comparable before the intervention. Specifically, we conducted our randomization 50 times and choose the assignment that minimized the difference in covariates between groups (this is known as the "minmax method", see Bruhn and McKenzie, 2009). These covariates included the baseline levels of our measures of child development (CREDI and CDI), the measure of a mother's play interactions with her child (FCI play subscale), the measure of maternal anxiety (GAD-7), the child's age and sex, the household asset index, and the number of call attempts needed to complete the baseline survey. By design, this method led to two comparable experimental groups (see Table 1).

2.5 Theory of change

In this section, we discuss what we expect to find for each of the research questions presented in section 1.2. Specifically, we tie together the hypotheses that we introduced in section 2.2.

The intervention seeks to address two separate but related challenges common to LMICs: low-income mothers lack information on how to support their children's development and, consequently, their sons and daughters often fail to reach their full developmental potential. We verified that these are indeed pressing needs among study participants by measuring mother-child interactions and children's overall and language development at baseline (for a description of these measures, see section 2.3).

The input offered by the intervention to address these challenges are 85 audio recordings delivered by phone over 18 weeks (for a description of the intervention, see section 2.4). We check that this input is being delivered as expected by tracking the number of calls made to each mother in the treatment group during the study, as mentioned in section 2.3.

The expected outputs are that mothers listen to the recordings and understand their content. For the former, we track the proportion of calls that mothers accept and the number of seconds that mothers stay on the phone before hanging up relative to the full call duration. For the latter, we track the share of correct touch-tone responses at the end of each call, as also mentioned in section 2.3 (for a description of these questions, see section 2.4).

The expected outcomes are that mothers update their knowledge and beliefs on child development and how to interact with their children, change such interactions, and feel more efficacious/less anxious as a result. We plan to check whether this is the case by measuring these outcomes at endline (for details on these measures, see section 2.3).

Lastly, the expected impact is that children improve their overall development and language development. We plan to measure both of these indicators once again at endline.[9]

2.6 Sample

We constructed a convenience sample of 2,433 mothers across 250 *anganwadi* centers across two blocks (Khatima and Jaspur) of one district (Udham Singh Nagar) of Uttarakhand.[10] In these centers, we recruited 2,433 mothers with at least one child (ages 6 to 30 months). The study's unit of analysis is the mother or her youngest child in that age range (depending on whether we focus on mother- or child-level outcomes).[11] We focus on the youngest eligible child in that age range because we expect to see larger effects among younger children and to keep the time costs of the survey manageable for mothers. Our statistical power calculations suggest that, even under conservative assumptions of 20% attrition, we should be able to detect small intent-to-treat effects of 0.095 standard deviations.[12]

2.7 Variations from the intended sample, and non-compliance

We do not expect attrition to exceed the (conservative) 20% that we have already factored into our statistical power calculations (see section 2.6), based on recent studies with similar

[9] It is possible (indeed, likely) that some of the mothers in the treatment group will not have listened to all recordings by the end of our study (e.g., mothers may miss calls, which would require *Dost* to call them back, extending the time it takes them to complete a module and delaying their exposure to the rest of the material). Yet, we want to know how much the average mother and child gained from the intended exposure.

[10] In India's Integrated Child Development Services (ICDS), "blocks" refer to administrative units that are one level below the district. There are 13 districts and 95 blocks in Uttarakhand.

[11] In our sample, there are 14 mothers with twins in that age range. In these cases, we focus on the child the mother named first during her baseline interview.

[12] These calculations are for a statistical power of 0.8, a statistical significance level of 0.05 (with two-sided tests), and an assumption that randomization strata fixed effects and baseline covariates explain 45 percent of an outcome's endline variation. They refer to the average intent-to-treat effect for the study's "global" measure of early childhood development (see section 3.3 for our discussion of main vs. secondary outcomes, and our approach to multiple hypothesis testing).

characteristics conducted by the Regional Office for South Asia of the Abdul Latif Jameel Poverty Action Lab (J-PAL SA), through which we are conducting our study.

We may encounter differential attrition at endline if mothers in the control group have changed their numbers and/or willingness to participate in the study. We will minimize this possibility, however, by reminding these mothers that they will receive the intervention after the study if they participate in the endline (see section 2.4.2). We discuss how we will address differential attrition in section 3.2.2).

We believe cross-over or "contamination" across experimental groups is highly unlikely. *Dost* chooses to whom it delivers audio recordings, and it will not do so for the control group until the study's endline data collection is complete.

Non-compliance with the intervention is also unlikely. In the first 10 weeks of the study, 86.1% of the mothers in the treatment group accepted at least one call per week, and they accepted 2.8 calls per week on average.[13]

2.8 Data collection and processing

We adhere to J-PAL SA's strict data collection procedures, including high-frequency checks for electronic forms, spot-checks and accompaniments, and weekly monitoring and debriefs for enumerators (see Glennerster, 2017; J-PAL, 2017).

3 Empirical analysis

3.1 Statistical model

3.1.1 Average intent-to-treat effects

We will estimate the intent-to-treat (ITT) effect of the offer of the intervention by fitting the following model:

$$Y_{ic}^{t=1} = \alpha_c + \beta^t T_{ic} + \boldsymbol{\delta}' \boldsymbol{X}_{ic}^{t=0} + \epsilon_{ic}^t \tag{1}$$

where Y_{ic}^t is the outcome of interest for mother or child i from center c at endline ($t = 1$), T_{ic} is an indicator variable for random assignment to the treatment group, and $\boldsymbol{X}_{ic}^{t=0}$ is a vector of baseline covariates at baseline ($t = 0$) selected through a LASSO procedure, from

[13] To complete all of the intervention's 85 calls over the study period, each mother needs to answer approximately 3.1 calls per week.

$Y_{ic}^{t=0}$ (whenever available), the child's age (in months) and sex, the mother's highest level of education, whether the mother is an adolescent, an index of household assets, and the number of call attempts needed to complete the baseline survey. The α_c parameters are early childhood center fixed effects (i.e., randomization strata fixed effects). The coefficient of interest is β^t, which captures the average causal effect of the intervention.

3.1.2 Heterogeneous intent-to-treat effects

We will also fit the following model to test for heterogeneous ITT effects:

$$Y_{ic}^{t} = \alpha_c + \beta^t T_{ic} + \theta^t T_{ic} * C_{ic}^{t=0} + \zeta^t C_{ic}^{t=0} + \boldsymbol{\delta'} \boldsymbol{X}_{ic}^{t=0} + \epsilon_{ic}^{t} \tag{2}$$

where $C_{ic}^{t=0}$ is a variable indicating that a child does not belong to the sub-group of interest (a child is not female; a child is not in the lowest quartile of $Y_{ic}^{t=0}$) and everything else is as defined above. Here, β^t captures the ITT effect for the sub-group of interest.[14] We will explore other dimensions of heterogeneity (e.g., by mother's household assets, whether she is an adolescent, and level of education) as exploratory.

We will complement the analysis of heterogeneous effects by child development specified above in two ways, following Ganimian et al. (2023). First, we will estimate quantile treatment effects using local polynomial regressions of endline scores on endline percentiles separately by experimental group. Then, we will estimate average treatment effects by baseline score using local polynomial regressions of endline scores on baseline percentiles separately by experimental group. We will plot the bootstrapped 95% confidence intervals.

3.1.3 Dose-response relationship

All results from the ITT estimates above will be based on the average number of calls that mothers receive. We will also present local average treatment effects (LATE) estimates of the impact of actually receiving and accepting the calls and of predicted treatment effects at different levels of program take-up. Specifically, we will estimate the dose-response relationship between the number of calls that mothers accept and children's overall and language development using:

$$Y_{ic}^{t} = \alpha_c + \mu^t A_{ic} + \gamma^t Y_{ic}^{t=0} + \boldsymbol{\delta'} \boldsymbol{X}_{ic}^{t=0} + \eta_{ic}^{t} \tag{3}$$

[14] We will deprioritize testing for effects among the more advantaged groups (e.g., being male) or for differences across groups (e.g., between female and male children).

where A_{ic} is the number of calls that mother i from center c has accepted (which is zero for all mothers in the control group), and everything else is defined as above. Since exposure may be endogenous to expected gains from the intervention, we instrument for the number of accepted calls with the randomized assignment to the program. Then, we will present the Average Causal Response (ACR) for mothers who accepted all 85 phone calls.

We will investigate whether the two assumptions required to estimate the ACR are met, following Muralidharan et al. (2019). We will conduct three analyses to examine whether there are constant treatment effects across children. First, we will examine whether the ITT effects are constant across the full distribution of baseline scores. Then, we will test whether we can reject the null hypothesis that the estimates from equation (3) and estimates using a value-added specification are equal.[15] Next, we will explore whether the constant term in the value-added specification (corresponding to zero accepted calls) is similar when using the full sample and when estimated using only the data in the treatment group. We will conduct two analyses to examine whether the relationship between the number of accepted calls and treatment effects is linear. First, we will plot value-added impact estimates against the number of accepted calls using only the treatment group. Then, we will model the dose-response relationship between the number of accepted calls and value-added quadratically to test whether we can reject the null hypothesis that this relationship is linear.

3.2 Statistical methods

3.2.1 Estimation

We will estimate equations (1) and (2) using ordinary least-squares (OLS) regressions and equation (3) using instrumental-variable (IV) regressions. It is common practice to cluster standard errors at the treatment level in randomized evaluations (cf. Abadie et al., 2022). Accordingly, we will not use clustered standard errors because mothers were individually randomized into experimental groups and we observe only one child per mother.[16]

We will use randomization inference (RI) to assess whether the re-randomization procedure led to unexpected consequences as a robustness check (Young, 2019). Specifically, we will replicate the same re-randomization procedure with 1,000 RI iterations (cf. Heß, 2017).

[15] Specifically, we will check whether the p-value from the difference-in-Sargan test of the equivalence of results is statistically insignificant.

[16] de Chaisemartin and Ramirez-Cuellar (2020) suggest making an exception for randomized trials with small randomization strata (such as pairwise randomized trials). Our study does not fall into this category of trials.

3.2.2 Rules for handling missing values

We expect to encounter two types of missing data: attrition (i.e., mothers not participating in the endline) or missing values (i.e., mothers participating in the endline, but not answering specific questions therein).

Missing values due to attrition. We will address the first type of missingness as follows. First, we will document the overall attrition rate. Then, we will investigate whether attrition is systematically related to intervention assignment by fitting a version of equation (1) that replaces the outcome variable with an indicator variable for not participating in the endline. Next, if we find differential attrition, following Barrera-Osorio et al. (2018) we will exploit tracking information and the number of calls needed to survey a respondent to model their propensity to attrit (Behaghel et al., 2014; Molina Millán and Macours, 2017).[17]

Missing values due to non-response. We will address the second type of missingness as follows. For missing responses on outcome variables, we will scale responses using item-response theory (IRT) models that account for missing values by using concurrent calibration via marginal maximum likelihood estimation (Kolen and Brennan, 2004), given that non-response on specific questions is akin the missingness in any non-equivalent anchor test (NEAT) design in which not all respondents are administered the same questions.[18] For missing responses on covariates, we will investigate the robustness of our results to the inclusion of such covariates and of the observations that are missing such covariates.

3.2.3 Definition and rules for handling outliers

We do not expect to encounter outliers because all of our outcome variables are measured on pre-determined scales (e.g., a mother may answer all or none of the questions in the GAD-7 affirmatively). Therefore, we will not seek to identify outliers or winsorize results.

3.3 Multiple hypothesis testing

We account for multiple hypothesis testing by clearly pre-specifying two measures of early childhood development (CREDI and CDI) as primary outcomes of interest. We will compare each outcome in the treatment and control group at endline (one comparison per outcome;

[17] This approach improves upon conventional inverse-probability weighting (IPW) and Lee (2009) bounds estimations.

[18] Following the CREDI scoring manual, if a mother refuses to answer at least five questions of an instrument, we mark the corresponding overall score as missing.

two comparisons in total). We will report both unadjusted p-values and p-values adjusted to control the false discovery rate (FDR).[19] Specifically, we will conduct the study's statistical tests in the following order.

The CREDI is a "global" measure of early childhood development that also captures a child's language development—therefore, in the group comparison for the CREDI, we do not apply an adjustment for multiple hypothesis testing. In contrast, the group comparison for the CDI is a second test, which warrants such an adjustment.

Next, we hypothesize mothers' interactions with their children serve as most important intermediate outcome and test for impacts on the FCI play subscale (test number three).

Our analyses of effects among subgroups will focus on two additional group comparisons (see section 3.1.2). Starting with the CREDI, we will assess program impacts among the quartile of less-developed children (test four) and impacts among girls (test five).

After that, we will explore (average) program impacts on mothers' beliefs (test six), anxiety levels (test seven), and self efficacy (test eight).

Lastly, we will repeat the above subgroup analyses for the CDI (tests nine and ten). We de-prioritize additional subgroup analyses and relegate them to exploratory analyses (e.g., group comparisons in the top quartile of baseline child development, among boys, etc.).

We prioritize the study's statistical tests in this order. For example, we will account for three tests in our analyses of impacts on the FCI play subscale, four tests in the group comparison of the CREDI among the quartile of less-developed children, five tests in the respective group comparison among girls, etc. As we discussed in section 3.1.3, we will estimate both ITT effects and LATE effects to obtain the ACR for mothers who accepted all calls from the intervention. We will not double-count these two types of analyses in our adjustments for multiple-hypothesis testing.

4 Limitations and challenges

To our knowledge, ours will be one of the first studies on the impact of an inexpensive, low-tech, and easily scalable approach to foster early psychosocial stimulation in an LMIC. As such, one of our primary objectives is to gauge the level of demand for this intervention. It seems possible that demand for the intervention declines over time due to multiple factors (e.g., not meeting mothers' expectations or changes in mothers' availability to receive calls).

[19] Multiple hypothesis testing and advancements over "basic" FDR methods (such as Storey's q) are an active area of research; we will choose a "modern" method such as Boca and Leek's FDR regression (see Korthauer et al., 2019).

If this were the case, we would leverage *Dost*'s call data to document changes in demand, identify the optimal timing (e.g., days of the week and time of the day) for higher take-up, and understand whether differences in take-up predict differences in mother/child impacts. This analysis would be helpful both to adjust this specific intervention and to inform the design of similar initiatives in LMICs (as Cortes et al., 2021, do in a high-income country).

We carefully selected instruments that have been previously administered in LMICs and have demonstrated adequate psychometric properties in such settings. Yet, it is possible that these scales are not sensitive to changes induced by the intervention. We will assess the extent to which this is an issue by analyzing the internal-consistency reliability and validity of these instruments administered over the phone, following Arteaga and Trias (2021). If we find consistently null results and funding allows, we may also conduct qualitative interviews with a sub-sample of mothers in the treatment group to try to understand what other changes may have occurred in mother-child interactions and children's development. This analysis could offer practical guidance how to measure the effect of similar programs.

We see this study as answering the initial question of whether a phone-based intervention can elicit sufficient take-up to improve the child-rearing practices of low-income mothers. This is why we decided to use self-reported measures of interactions and child development, rather than more costly and time consuming measures of overall and language development. It is possible that we observe impacts on take-up and interactions, but not on development. If this were the case, we would conduct a second study in which we would employ multiple measures of overall and language development, including observations and assessments. This strategy avoids the possibility of deploying an involved child measurement strategy only to find not enough mothers stay with the program or that their practices do not change. Given our relationship with *Dost* and our conversations with donors, we are well positioned to undertake a follow-up study focused on child development outcomes at a larger scale.

Tables

Table 1: *Balancing checks between experimental groups*

	N	(1) Control Mean/SD	N	(2) Treatment Mean/SD	(3) Difference (1)-(2)
Panel A: Child outcomes					
Child development (CREDI)	1214	0.000	1219	-0.011	0.011
		[1.000]		[0.974]	(0.040)
Vocabulary score (CDI)	1214	0.000	1219	-0.024	0.024
		[1.000]		[0.969]	(0.040)
Panel B: Maternal outcomes					
Family care index (FCI), play scale	1214	0.000	1219	0.012	-0.012
		[1.000]		[1.000]	(0.040)
Anxiety score (GAD-7)	1213	0.000	1219	-0.015	0.015
		[1.000]		[0.991]	(0.040)
Panel C: Background characteristics					
Child is female	1214	0.498	1219	0.489	0.009
		[0.500]		[0.500]	(0.020)
Child age (in months)	1214	18.023	1219	17.808	0.215
		[6.894]		[6.762]	(0.279)
Asset index	1214	0.000	1219	0.027	-0.027
		[1.000]		[1.035]	(0.040)
Call attempts needed to complete the survey	1214	2.452	1219	2.490	-0.038
		[1.727]		[1.752]	(0.068)

Notes. This table compares individuals in the control and treatment groups at baseline. It shows the mean and corresponding standard deviations for each variable (in brackets), and it compares both experimental groups, including randomization-strata fixed effects, showing the mean difference and corresponding standard errors (in parentheses). Except for child age and the number of call attempts needed to complete the survey, continuous variables are standardized and centered with respect to the control group. CREDI scores are aggregated following the instrument's official scoring guidelines. CDI scores reflect the total vocabulary score. FCI scores are aggregated using item response theory (IRT) with a two-parameter logistic (2PL) model. GAD-7 scores are aggregated using IRT with a generalized partial credit model (PCM). The asset index reflects an inverse-covariance-weighted (ICW) average across eight yes/no questions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical levels.

References

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M., 2022. When Should You Adjust Standard Errors for Clustering? The Quarterly Journal of Economics 138, 1–35. doi:10.1093/qje/qjac038.

Andrew, A., Attanasio, O.P., Augsburg, B., Day, M., Grantham-McGregor, S.M., Meghir, C., Mehrin, F., Pahwa, S., Rubio-Codina, M., 2020. Effects of a scalable home-visiting intervention on child development in slums of urban India: Evidence from a randomised controlled trial. Journal of Child Psychology and Psychiatry 61, 644–652.

Arteaga, I., Trias, J., 2021. Can technology narrow the early childhood stimulation gap in rural Guatemala? Results from an experimental approach. *Unpublished manuscript*. Washington, DC: The World Bank.

ASER, 2021. Annual status of education report (rural) 2021. New Delhi, India: ASER Centre.

Attanasio, O.P., Baker-Henningham, H., Bernal, R., Meghir, C., Pineda, D., Rubio-Codina, M., 2018. Early stimulation and nutrition: The impacts of a scalable intervention. (NBER Working Paper No. 25059). Cambridge, MA: National Bureau of Economic Research (NBER).

Attanasio, O.P., Fernández, C., Fitzsimons, E.O.A., Grantham-McGregor, S.M., 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. BMJ 349, g5785.

Babikako, H.M., Bourdon, C., Mbale, E., Aber, P., Birabwa, A., Chimoyo, J., Voskuijl, W., Kazi, Z., Massara, P., Mukisa, J., et al., 2022. Neurodevelopment and recovery from wasting. Pediatrics 150.

Banerjee, A., Chassang, S., Montero, S., Snowberg, E., 2020. A Theory of Experimenters: Robustness, Randomization, and Balance. American Economic Review 110, 1206–1230. URL: https://pubs.aeaweb.org/doi/10.1257/aer.20171634, doi:10.1257/aer.20171634.

Barrera-Osorio, F., de Barros, A., Filmer, D.P., 2018. Long-term impacts of alternative approaches to increase schooling: Experimental evidence from a scholarship program in Cambodia. Working Paper WPS8566. The World Bank. Washington, D.C. URL: http://documents.worldbank.org/curated/en/838871535033752683/Long-term-impacts-of-alternative-approaches-to-increase-schooling-evidence-fro

Beam, E., Mukherjee, P., Navarro-Sola, L., 2022. Lowering barriers to remote education: Experimental impacts on parental responses and learning. (IZA Discussion Paper No. 15596). Bonn, Germany: Institute for the Study of Labor (IZA).

Behaghel, L., Crépon, B., Gurgand, M., Le Barbanchon, T., 2014. Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models. The Review of Economics and Statistics 97, 1070–1080. URL: https://doi.org/10.1162/REST_a_00497, doi:10.1162/REST_a_00497.

Berlinski, S., Galiani, S., Gertler, P., 2009. The effect of pre-primary education on primary school performance. Journal of public Economics 93, 219–234.

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

Blimpo, M.P., Carneiro, P., Jervis, P., Pugatch, T., 2022. Improving access and quality in early childhood development programs: Experimental evidence from The Gambia. Economic Development and Cultural Change 70, 1479–1529.

Bouguen, A., Filmer, D., Macours, K., Nadeau, S., 2014. Preschool and parental response in a second best world: Evidence from a school construction experiment. The Journal of Human Resources 53, 474–512.

Bruhn, M., McKenzie, D., 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. American Economic Journal: Applied Economics 1, 200–232. doi:10.1257/app.1.4.200.

CDC, 2011. Building adult capabilities to improve child outcomes: A theory of change. Cambridge, MA: Center on the Developing Child at Harvard University. URL: https://developingchild.harvard.edu/resources/building-adult-capabilities-to-improve-child-outcomes-a-theory-of-change/.

de Chaisemartin, C., Ramirez-Cuellar, J., 2020. At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata? arXiv:1906.00288 [econ] URL: http://arxiv.org/abs/1906.00288. arXiv: 1906.00288.

Chang-Lopez, S., Walker, S., Grantham-McGregor, S., Powell, C., 2020. Parent manual: Activities for children up to age 3 years. Kingson, Jamaica: Caribbean Institute for Health Research, The University of West Indies.

CISCD, 2000. From neurons to neighborhoods: the science of child development. Committee on Integrating the Science of Child Development. Washington DC: National Academy Press.

Cortes, K.E., Fricke, H., Loeb, S., Song, D.S., York, B.N., 2021. Too little or too much? actionable advice in an early-childhood text messaging experiment. Education Finance and Policy 16, 209–232.

De Man, J., Absetz, P., Sathish, T., Desloge, A., Haregu, T., Oldenburg, B., Johnson, L.C., Thankappan, K.R., Williams, E.D., 2021. Are the PHQ-9 and GAD-7 suitable for use in India? a psychometric analysis. Frontiers in psychology 12, 676398.

Dean, J., Jayachandran, S., 2019. The impact of early childhood education on child development in rural India. *Unpublished manuscript*. Karnataka, India: Abdul Latif Jameel Poverty Action Lab (J-PAL).

Doss, C., Fahle, E.M., Loeb, S., York, B.N., 2019. More than just a nudge: Supporting kindergarten parents with personalized and differentiated text messages. Journal of Human Resources 56, 567–603.

Engle, P.L., Black, M.M., Behrman, J.R., Cabral de Melho, M., Gertler, P.J., Kapiriri, L., Martorell, R., Young, M.E., the International Child Development Group, 2007. Strategies to avoid the loss of developmental potential in more than 200 million children in the developing world. The Lancet 369, 229–242.

Fenson, L., Pethick, S., Renda, C., Cox, J.L., Dale, P.S., Reznick, J.S., 2000. Short-form versions of the macarthur communicative development inventories. Applied psycholinguistics 21, 95–116.

Ferdowshi, N., Imran, M.A., Trishna, T.A., 2021. Adaptation of the tool to measure parenting self-efficacy (TOPSE) in Bangladesh. Dhaka University Journal of Biological Sciences 30, 169–177. doi:10.3329/dujbs.v30i2.54643.

Floccia, C., Sambrook, T., Delle Luche, C., Kwok, R., Goslin, J., White, L., Cattani, A., Sullivan, E., Abbot-Smith, K., Krott, A., et al., 2018. Vocabulary of 2-year-olds learning english and an additional language: norms and effects of linguistic distance. v: General discussion. Monographs of the Society for Research in Child Development 83, 68–80.

Ganimian, A.J., Muralidharan, K., Walters, C.R., 2023. Improving early-childhood human development: Experimental evidence from India. Journal of Political Economy *Unpublished manuscript*. New York, NY: New York University (NYU).

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M., Grantham-McGregor, S.M., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. Science 344, 998–1001.

Glennerster, R., 2017. The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 175–243. doi:10.1016/bs.hefe.2016.10.002.

Grantham-McGregor, S.M., Adya, A., Attanasio, O.P., Augsburg, B., Behrman, J.R., Caeyers, B., Day, M., Jervis, P., Kochar, R., Makkar, P., Meghir, M., Phimister, A., Rubio-Codina, M., Vats, K., 2020. Group sessions or home visits for early childhood development in India: A cluster RCT. Pediatrics 146.

Grantham-McGregor, S.M., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L., Strupp, B., Group, I.C.D.S., et al., 2007. Developmental potential in the first 5 years for children in developing countries. The lancet 369, 60–70.

Grantham-McGregor, S.M., Powell, C.A., Walker, S.P., Himes, J.H., 1991. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: The Jamaican study. The Lancet 338, 1–5.

Hamadani, J.D., Tofail, F., Hilaly, A., Huda, S.N., Engle, P., Grantham-McGregor, S.M., 2010. Use of family care indicators and their relationship with child development in bangladesh. Journal of health, population, and nutrition 28, 23.

Heß, S., 2017. Randomization Inference with Stata: A Guide and Software. The Stata Journal: Promoting communications on statistics and Stata 17, 630–651. doi:10.1177/1536867X1701700306.

J-PAL, 2017. J-PAL Research Protocols. URL: https://drive.google.com/file/d/0B97AuBEZpZ9zZDZZbV9abllqSFk/view.

Jackson-Maldonado, D., Marchman, V.A., Fernald, L.C., 2013. Short-form versions of the spanish macarthur–bates communicative development inventories. Applied Psycholinguistics 34, 837–868.

Karuppannan, A., Ramamoorthy, T., Rammamoorthi, A., Ravichandran, L., 2020. Mother's knowledge on child's developmental milestones and parenting skills in kanchipuram district, tamilnadu: a descriptive cross sectional study. Int J Health Sci Res [Internet] 10, 242–7.

Kendall, S., Bloomfield, L., 2005. Developing and validating a tool to measure parenting self-efficacy. Journal of advanced nursing 51, 174–181.

Kern, S., 2007. Lexicon development in french-speaking infants. First Language 27, 227–250.

Knauer, H.A., Kagawa, R.M., Garcia-Guerra, A., Schnaas, L., Neufeld, L.M., Fernald, L.C., 2016. Pathways to improved development for children living in poverty: A randomized effectiveness trial in rural mexico. International Journal of Behavioral Development 40, 492–499.

Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking. 3rd ed., Springer, New York, NY.

Kopper, S., Sautmann, A., 2020. Best practices for conducting phone surveys. Abdul Latif Jameel Poverty Action Lab (J-PAL) 20.

Korthauer, K., Kimes, P.K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E.J., Hicks, S.C., 2019. A practical guide to methods controlling false discoveries in computational biology. Genome Biology 20, 118. doi:10.1186/s13059-019-1716-1.

Landry, S.H., Smith, K.E., Swank, P.R., 2006. Responsive parenting: Establishing early foundations for social, communication, and independent problem-solving skills. Developmental Psychology 42, 627–642.

Lee, D.S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies 76, 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x.

List, J.A., Pernaudet, J., Suskind, D.L., 2021. Shifting parental beliefs about child development to foster parental investments and improve school readiness outcomes. Nature communications 12, 5765.

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., Herzberg, P.Y., 2008. Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. Medical care , 266–274.

Lu, C., Black, M.M., Richter, L.M., 2016. Risk of poor development in young children in low-income and middle-income countries: An estimation and analysis at the global, regional, and country level. The Lancet 4, e916–e922.

Luoto, J.E., Garcia, I.L., Aboud, F.E., Singla, D.R., Fernald, L.C., Pitchik, H.O., Saya, U.Y., Otieno, R., Alu, E., 2021. Group-based parenting interventions to promote child development in rural kenya: a multi-arm, cluster-randomised community effectiveness trial. The Lancet Global Health 9, e309–e319.

MacPhee, D., 1981. Knowledge of infant development inventory: Manual. Chapel Hill, NC: Department of Psychology, University of North Carolina .

Martinez, S., Naudeau, S., Pereira, V., 2017. Preschool and child development under extreme poverty: Evidence from a randomized experiment in rural Mozambique. (Policy Research Working Paper No. 8290). Washington, DC: The World Bank.

McCoy, D.C., Sudfeld, C.R., Bellinger, D.C., Muhihi, A., Ashery, G., Weary, T.E., Fawzi, W., Fink, G., 2017. Development and validation of an early childhood development scale for use in low-resourced settings. Population health metrics 15, 1–18.

McCoy, D.C., Waldman, M., Team, C.F., Fink, G., 2018. Measuring early childhood development at a global scale: Evidence from the caregiver-reported early development instruments. Early childhood research quarterly 45, 58–68.

Molina Millán, T., Macours, K., 2017. Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias. Discussion Paper 10711. IZA Institute of Labor Economics. Bonn. URL: http://ftp.iza.org/dp10711.pdf.

Muralidharan, K., Singh, A., Ganimian, A.J., 2019. Disrupting education? Experimental evidence on technology-aided instruction in India. American Economic Review 109, 1–35.

NCERT, 2019. Theme-based early childhood care and education programme. New Delhi, India: National Council of Educational Research and Training (NCERT).

Psacharopoulos, G., Patrinos, H.A., 2004. Returns to investment in education: a further update. Education Economics 12, 111–134.

Tofail, F., Hamadani, J.D., Mehrin, F., Ridout, D.A., Huda, S.N., Grantham-McGregor, S.M., 2013. Psychosocial stimulation benefits development in nonanemic children but not in anemic, iron-deficient children. The Journal of nutrition 143, 885–893.

UNICEF, 2017. UNICEF's programme guidance for early childhood development. New York, NY: United Nations Children's Education Fund (UNICEF).

Waldman, M., McCoy, D.C., Seiden, J., Cuartas, J., Fink, G., 2021. Validation of motor, cognitive, language, and socio-emotional subscales using the Caregiver Reported Early Development Instruments: An application of multidimensional item factor analysis. International Journal of Behavioral Development 45, 368–377. doi:10.1177/01650254211005560.

Walker, S.P., Chang, S.M., Powell, C.A., Grantham-McGregor, S.M., 2005. Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and

education in growth-stunted jamaican children: prospective cohort study. The lancet 366, 1804–1807.

Walker, S.P., Wachs, T.D., Gardner, J.M., Lozoff, B., Wasserman, G.A., Pollitt, E., Carter, J.A., Group, I.C.D.S., et al., 2007. Child development: Risk factors for adverse outcomes in developing countries. The Lancet 369, 145–157.

York, B.N., Loeb, S., Doss, C., 2019. One step at a time: The effects of an early literacy text-messaging program for parents of preschoolers. Journal of Human Resources 54, 537–566.

Young, A., 2019. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results. The Quarterly Journal of Economics 134, 557–598. doi:10.1093/qje/qjy029.

Young, M., 2002. From early child development to human development. Washington, DC: The World Bank.

Yousafzai, A.K., Rasheed, M.A., Rizvi, A., Armstrong, R., Bhutta, Z.A., 2014. Effect of integrated responsive stimulation and nutrition interventions in the Lady Health Worker programme in Pakistan on child development, growth, and health outcomes: A cluster-randomised factorial effectiveness trial. The Lancet 384, 1282–1293.