Early Childhood Development in the Slums of Cuttack, Odisha, India

Pre-analysis plan

23rd November 2015

1. Introduction

The very earliest years of childhood are key to fulfilled, productive and meaningful lives. Children's brain and physical development is at its most rapid during this time as they develop skills and capabilities that affect lifetime outcomes as diverse as earnings, wellbeing and criminality. Gaps that open up between children, often along familiar lines of wealth and income, during this stage typically persist and are exacerbated over time. Thus, these years are key to understanding the transmission of poverty across generations. For many children growing up in poorer countries, these earliest years don't offer conditions that are sufficient to reach their developmental potential. Poverty, malnutrition and disease-ridden and un-stimulating environments can all contribute to children falling short developmentally of what they otherwise would have been capable of. Excitingly, however, a vibrant research agenda demonstrates that a child's development is not predetermined but highly malleable: it is heavily affected by environmental factors which can be altered by policy or behaviour change. This creates a clear rationale for intervening early in life, especially for the most disadvantaged children.

In this study researchers at the IFS, UCL and Yale, in collaboration with Pratham, JPal-SA, will evaluate an intensive Early Childhood Development (ECD) intervention, in Cuttack, India. The home visiting programme aims to increase the development of cognitive and language skills, through increasing the level of psychosocial stimulation children are exposed to. It consists of 18 months of weekly home visits, from trained local women following a structured curriculum, and activities that mothers are encouraged to do between visits. The aim is to improve levels of interaction and attachment between mothers and their infants, creating a more stimulating environment for the child and increasing his or her expected level of development.

The impact evaluation of this intervention, which is based on a cluster randomised control trial design, will directly and rigorously study the effectiveness of the home visiting intervention over a range of child development indicators. Further work will attempt to analyse the mechanisms through which the programme impacted (if at all) these measures. Our main hypotheses of interest are:

1. the home visiting intervention will have a positive average effect on children's development in four domains: (i) cognitive development, (ii) receptive language development, (iii) expressive language development and (iii) fine motor development.
2. these effects on child development will be driven (at least partially) through (i) increased quality of the home stimulation environment, (ii) increased maternal time spent on high stimulation activities with children and (iii) increased maternal knowledge of child development.

The evaluation design includes a baseline survey before the intervention as well as an endline survey after the implementation of the program.

This pre-analysis plan sets out, ex-ante, how we will analyse data from the endline survey and child development assessments, in combination with baseline data, to assess whether the home-visiting intervention had a significant impact on various child development outcomes. We describe how we plan to create our various outcomes of interest (both final and intermediate) and our empirical strategy for estimating the impact of the home-visiting intervention on child development outcomes and for assessing the statistical significance of such estimates. We also include our methods for dealing with multiple hypothesis testing.

2. Research strategy

2.1. Sampling and randomisation

For a detailed description of our sampling and randomisation strategy please consult our baseline report (Andrew et al., 2015). Here follows a summary.

Our study follows a cluster randomised controlled trial design. Clusters are sahis (or slums) in Cuttack, India of which we selected 54 to include in our study. We selected these to ensure there were sufficient children in our target age range.

Within each cluster we aimed to approach 9 children for enrolment in our study (accounting for an anticipated positive refusal rate and clusters with fewer than 9 eligible children). In each of the 54 clusters we completed a census to identify all the children who met our inclusion criteria (aged 10-20 months at baseline, living in the 54 study sahis, excluding twins and children with physical or mental disabilities). If we found fewer than 9 eligible children in a cluster then we aimed to include all of them in our study while if we found more than 9 we randomly selected 9 whom we aimed to include. This gave us an initial list of 459 children who we attempted to collect baseline data from. In some cases (around one fifth) baseline could not be completed due to refusal, child being out of age range or the household having (or planning to) relocated. These children were then excluded from the study and, in clusters with additional children, an additional eligible child was randomly drawn to replace them. Overall, this sampling method gave us a total of 421 target children for whom we have complete baseline data. This process is summarised below in Figure 1.

In our sample eligibility for the home-visiting intervention was randomised at the cluster (sahi) level. We randomly selected 27 sahis to receive the home-visiting intervention while the other 27 formed the control group. We stratified our randomisation of treatment status on the number of children in the target age range in the sahi. There were two stratas: (1) sahis that had more than nine children in the target age range (10-20 months) and therefore only some of the children in the sahi would be eligible for the intervention, a total of 36 sahis, and (2) sahis that had fewer children in the target age range where all the children in the sahi would be eligible for the intervention, a total of 18 sahis. We stratified on this indicator to increase balance in the case where the effectiveness of the intervention differed depending on whether all or only some of the children in the sahi were eligible.

For endline data collection we aim to collect data on all those 421 target children who form our baseline sample. In Section 3.1 we outline how we test for differential attrition by treatment group which could bias our simple estimates of the effect of the home visiting intervention.

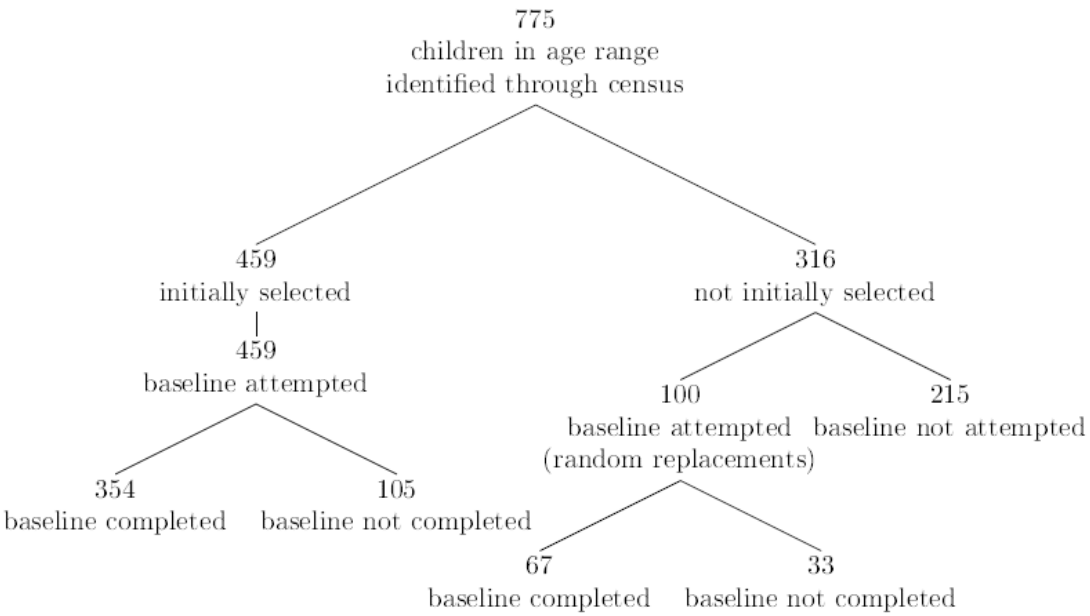

Figure 1: Process of selecting target children

2.2. Fieldwork

2.2.1. Instruments

In the evaluation of the effects of the home visiting intervention we will draw on three key sources of data: (i) Bayley Scales of Infant Development III, (ii) household survey data and (iii) administrative data on the operation of home visits (frequency, person delivering, and indicators of quality). In addition we will use our baseline data, including our baseline measure of child development – the Ages and Stages Questionnaires, third edition (ASQ-3) – to increase the precision of our estimates.

Bayley Scales of Infant Development III: The Bayley Scales of Infant and Toddler Development (Bayley-III) are seen internationally as a gold standard in measuring child development. The Bayley-III takes 45-60 minutes to administer and must be administered by a psychologist or other professional with expertise in assessing children.

The Bayley-III measures cognitive development, receptive communication, expressive communication, fine motor development, gross motor development, socio-emotional development and various components of adaptive behaviour for children 0-42 months of age. Each scale is formed of individual items that progressively indicate increasing levels of development. On each item the child can either score a 1 if he or she met the scoring criteria for that item or 0 otherwise. The starting point for each child on each scale is determined by his or her age. The tester begins by administering the first 3 items, if the child scores a 1 for each then the assessment continues forwards. If the child scores a 0 on any of the first three consecutive items then the tester goes back to the starting point for the previous age and begins again there. Items are then administered in order until

the child has obtained 0s on five consecutive items. At this point the assessment of that scale stops.

We used the Bayley-III to measure of five key domains of child development – cognitive development, receptive communication, expressive communication, fine motor development and gross motor development. We used 8 trainers who had backgrounds in psychology, physiotherapy and social work. The team of testers received training on the concepts behind psychometric testing in general and the Bayley-III in particular in addition to extensive training on the administration and scoring of each item. The training also involved many practice administrations of the test. Overall the training lasted for around 30 days and each tester practiced on a minimum of 10 children (10-15 children). The Bayley assessments took place in designated testing centres within the communities. We endeavoured to ensure the testing centres met the following criteria but in some cases this may not have been possible:

- Quiet, free of distraction. For example minimal things or objects around, displayed on wall. Restricted frequent movement in and out of the door.
- Safe and clean place- no sharp edges or open electric points.
- Comfortable seating arrangement.
- Adequate lighting and ventilation.

Household Survey: The household survey contained the following 7 modules:

- 1 – Dwelling and the household
- 2 – Workforce
- 2A – Roster
- 3 – Children younger than 16 years (Childcare, education and health)
- 4 – Biological mother
- 4A – Depressive symptoms (biological mother)
- 5 – Main caregiver (if different from mother)

The survey was administered by 14 female surveyors who received approximately seven days of training. The surveyors were organised into seven teams of two with the modules split between them. There were three supervisors and a field manager to monitor quality. These staff sat in on 10% of surveys and conducted random spot checks on another 10%. They also kept track of progress using the tracking sheets.

Administrative Data: Throughout the intervention we have been collecting administration data which recorded when each visit took place, who took the visit and who was the caregiver with the child, whether it was supervised, which material was covered and the home visitor's (and supervisor's where appropriate) assessment of the visit. Using this data we will construct the following variables to descriptively analyse the extent to which the intervention was delivered as we had planned in terms of the logistics of the visits:

- number of visits

- time lag in between visits
- consistency of visitor (document visitor turn-over)
- how many visits were supervised
- frequency of supervision
- consistency of caregiver
- presence of child in visit

Since our empirical strategy for assessing the home visiting intervention is an intent to treat analysis we will not use these measures of intensity of the intervention in our analysis.

2.2.2. Data Collection

Training for the Bayley-III happened first (19th March-1st May) and Bayley-III data collection began on the 4th May. Training for the household survey followed (14th May – 23rd May) and data collection began on the 25th May. All field staff involved in data collection (Bayley and household survey) were blind to treatment allocation. In cases where the household was reluctant to allow data to be collected (particularly taking the child to the test centre for the Bayley) the home visitors or other personnel from the intervention were never engaged to encourage the household since we worried this would lead to differential attrition between treatment groups as well as unblinding the data collection stage.

2.2.3. Data Processing

All data processing and cleaning will all be done prior to merging in treatment assignment. This means that decisions on any unanticipated issues that arise will be made without any knowledge of the treatment assignment of different units.

All data processing will be conducted on Stata 13 using do-files. This means we will have a record of every decision made in processing and cleaning.

Household questionnaire

Processing and cleaning of the household questionnaire will involve:

- Checking all ID variables using names and baseline data. Correcting where necessary.
- Checking all age and date of birth data with baseline. Where there are discrepancies we will use the cleaned date of birth from baseline since these were verified.
- Checking all skip patterns were followed correctly and altering any responses accordingly
- Checking for extreme values in continuous variables that are clearly recording errors and changing to missing accordingly

- Checking for obvious unit errors (e.g. birth weight in kgs rather than grams) and correcting accordingly
- Check for clear inconsistencies between question answers and, if obvious, correcting accordingly

Bayley-III scoring and standardising

We will check the assessment was administered as follows:

- Checking starting points were chosen correctly, in line with the age of the child, and modifying data where possible to correct any mistakes
- Checking starting rules were carried out (going back to previous starting point if the child got any 0s in the first three consecutive questions)
- Checking finishing rules were carried out (stopping after the first set of five consecutive questions where the child scored 0 for each). Where possible (if test continued past where it should have stopped), modify the data accordingly.
- Checking for missing items

Our main outcomes measures will be internally standardised Bayley-III for cognition, receptive language, expressive language, and fine motor. We will perform this standardisation relative to the control group since the control group [1] should be unaffected by the interventions and therefore this allows us to see the effect of the interventions relative to an unaffected population. Our procedure will be as follows:

For each scale,

- we will remove tester effects by running a regression of the total raw score on tester dummies
- we will use STATA's lowess command to non-parametrically estimate the evolution of the mean and variance of the residuals of the raw scores (i.e net of tester effects) for those children in the control group with age (in days)
- we will subtract our estimate of the age-specific mean (of the control group) from each residual of the raw score and divide by the age-specific standard deviation (of the control group)

This procedure will leave us with a z-score for each scale of the Bayley and we can then interpret the magnitude of effect sizes relative to one standard deviation of the control group.

In addition, for robustness and comparability with other literature, we will also construct the externally standardised scores for each domain using conversion tables provided in the test manual . We will also report the effect on raw scores controlling for polynomials in age as an alternative means of removing the age effect. In these specifications we will control for tester effects by including dummy variables for each tester in the regression.

[1] Since we plan to clean and standardise all data prior to merging in treatment status we will create a simulated treatment variable (by randomly allocating 27 clusters to each treatment group) during data cleaning and processing. This will bear no relation to actual treatment status. This will allow us to write all code to construct standardised scores, to run regressions etc. that requires the treatment variable. Once we have finished this code we will replace this simulated treatment variable with the true treatment variable.

3. Empirical Strategy

3.1. Attrition Analysis

Before we estimate any impacts of the home visiting program on outcomes we will first consider the attrition rate and balance of our endline sample. Since our methodology is a randomised controlled trial our estimates rely on the assumption that randomisation created a treatment and control group that were (in expectation) identical at baseline so that any differences we observe in key outcomes at follow-up can be attributed to the effect of the home visiting intervention. As discussed in detail in our baseline report (Andrew et al. 2015) overall we found that randomisation had successfully created two treatment groups that appeared the same over most background characteristics and factors we would consider key to child development[2]. However, we did identify imbalances in some baseline characteristics that may be related to child development[3], most notably wealth as measured by household asset ownership. Overall, however, we see our two treatment groups as being suitably balanced over the most important determinants of child development and thus our empirical strategy will be well suited to estimating the impact of the intervention. However, this could be compromised if attrition (losing sample children between baseline and follow-up) is correlated to both treatment status and other characteristics that affect child development since this violates the assumption that the two groups are, in expectation, identical other than their treatment assignment.

We will assess attrition in our study in two ways. First, we will test whether attrition, A_{ij}, is significantly related to treatment, T_j, status using the following logit model:

$$A_{ij} = \beta T_j + \gamma X_{ij} + \varepsilon_{ij}$$

We will estimate the parameters of this logit model both with and without controlling for key household and child baseline characteristics X_{ij} such as age, gender, wealth index, child developmental level etc. We will test the null hypothesis that $\beta = 0$, accounting for clustering of errors (ε_{ij}) in our analysis. If β is not significantly different from zero this implies that there is no evidence that attrition was related to treatment status and we will not adjust any estimates for differential attrition. If we reject the hypothesis that $\beta = 0$ this implies that there is differential attrition between treatment groups. It is important to note that if we find some of the baseline characteristics in the above regression are significantly different from zero (jointly or singly) but β is not then this will not be evidence that differential attrition will bias our estimates. However, if attrition is strongly correlated with baseline characteristics this may impact the generalisability of our findings.

Second, we will compare the baseline characteristics of the treatment and control groups for the attriters and non-attriters separately. We will do this by simply comparing the means and testing the hypothesis that the characteristics of those children that attrition (and

[2] Household size and structure; Religion; Dwelling characteristics; ASQ-3 – problem solving, communication, fine motor and gross motor; All anthropometric measurements; Morbidity; Quality of home environment in terms of play materials and play activities; Sanitation; Food preparation environment; Birth and breastfeeding; Diet and dietary diversity; Knowledge of child development; Time use; Depressive symptoms; Maternal empowerment; Maternal employment, wages and earnings; Paternal employment; Non-work income and transfers; Household expenditures; Household savings; Physical dwelling characteristics; Facilities in sahi; Social programmes in sahi

[3] Personal social development as measured by ASQ-3; Number of books and magazines in household; Maternal education (% completing fifth standard or higher); Maternal functional literacy; Paternal wages and earnings; Asset ownership; Household debt

correspondingly those that remain in the sample) are balanced across treatment groups. In this analysis we will control for clustering and multiple hypothesis testing.

If we do find significant evidence of differential attrition across treatment groups we will explicitly model the attrition (based on observable characteristics at baseline) and report estimates that are corrected for it, reporting bounds on our estimates where necessary.

3.2. Empirical specification

Our main analysis will be an intent-to-treat evaluation of the effect of the home visiting programme on those children who were targeted for the intervention. That is to say we are estimating the impact of being in a sahi that was allocated to the treatment group, and thus that the child was eligible for the home visiting programme. This may be different from the impact of actually receiving the intervention if some households decide not to participate in the programme even though they were eligible, for example if they perceived the programme would be of no benefit to their child.

For each continuous outcome of interest (final or intermediate) we will estimate the impact of eligibility for the home visiting programme by estimating, by Ordinary Least Squares, the parameters of the following linear regression model:

$$y_{ij} = \beta T_j + \gamma X_{ij} + \varepsilon_{ij}$$

where y_{ij} is the outcome of interest for the child (or household) *i* in sahi *j*, T_j is a dummy variable equal to 1 if sahi *j* was allocated to the treatment group receiving the home visiting intervention and equal to zero otherwise. X_{ij} is our vector of control variables collected at the time of baseline, including a constant. In section 3.3 we describe exactly how we will determine which variables will enter X_{ij} for each outcome variable. ε_{ij} is a random error term which is allowed to be correlated at the cluster level but assumed to be independently distributed between clusters.

In the case of a binary outcome variable we will use a logistic regression model with the same treatment indicator and control variables as above.

In this regression framework the most interesting parameter is β, our estimate of the impact of being eligible for the home visiting programme. It is the size and significance of this parameter that will tell us the impact of the intervention on the outcome of interest and the degree of uncertainty associated with that estimate.

3.3. Control Variables

The purpose of including control variables, X_{ij}, in the regression model is to increase precision of our estimate. For a variable to be a valid and useful control variable it must explain some part of the variation in the outcome in question, y_{ij}, and be uncorrelated with the explanatory variable of interest, treatment status T_j. Therefore in a RCT set-up with baseline data ideal controls are baseline variables that are closely correlated with the outcome of interest. Since allocation to treatment group is independent of baseline characteristics all baseline variables will be independent of treatment status.

Unless stated otherwise we will always control for the baseline level of all outcomes related to that hypothesis since we expect strong time persistence in all outcomes of interest. When

we are estimating the impact on an indexed measures we will control for the baseline value of each of the components separately. In addition to these baseline measures of the outcome of interest we will also include a set of 'core' controls depending on whether the outcome in question is measured at the child, household or mother level:

- **Core child level controls:** Age in days, gender, parity (dummy for first child), mother's education (in years) at baseline
- **Core mother level controls:** Number of biological children at baseline, age, education (in years) at baseline
- **Core household level controls:** Joint household at baseline (binary variable), household size at baseline

3.4. Treatment of Outliers

We will drop children from our analysis with developmental outcomes (age standardised Bayley-III scores) lower than three standard deviations (estimated from our control group) below the mean (again estimated from our control group), i.e. those children with z-scores <-3. We do this because we see this as a sign of potential disability or that children performing very poorly on the assessment for other reasons, perhaps due to illness.

3.5. Missing data

We will not impute the values for any dependent variable (final or intermediate outcomes) at follow-up. Regarding missing data on control variables, we will check whether item non-response is correlated with treatment status. If it is not correlated, we will impute the missing covariate value with the average of the non-missing observations and this imputation will be accounted for with a dummy variable (we will check the robustness of our results by also estimating the regression without that covariate). If non-response in the baseline/covariate is correlated with treatment status, we will not use that covariate when estimating the regressions. In cases in which the percentage of observations with covariate missing data is less than 2%, we will simply work with the sample with non-missing data.

3.6. Adjusting Standard Errors

Note that the aggregate error, $v_j + \varepsilon_{ij}$, cannot be assumed to be independent between households (or children) since households living in the same sahis may be subject to correlated unobserved shocks or their unobserved characteristics may be correlated. Therefore for our inference we will cluster errors at the level of the sahi, allowing for arbitrary correlation between error terms of households in the same sahi.

3.7. One- and Two-tailed Hypothesis Tests

Unless otherwise specified we will use two-tailed hypothesis tests in determining the significance of all estimated effects of the intervention. However, as discussed in section 4.1 for our primary outcomes, child development as measured by scales of the Bayley-III, we will use one-tailed tests. We do this as we have a strong prior that the intervention will not harm the developmental levels of the children.

3.8. Multiple Hypothesis Testing

We plan to deal with multiple hypothesis testing in two ways. Firstly, within each group of outcomes (e.g. child developmental outcomes) we plan to aggregate outcomes into a single index using factor analysis (following the procedure described in the following section). We will then test whether the estimated effect on this aggregated index is significantly different from zero (or significantly greater than zero in the case of a one tailed hypothesis test). Secondly, we will also test each individual outcome measure within the group of outcomes and adjust the p-values for multiple hypothesis testing whether the estimated effect on each outcome is significantly different from zero (or significantly greater than zero in the case of a one tailed hypothesis test). We will use the Romano-Wolf step-down procedure (Romano and Wolf 2005a,b) for this adjustment. In sections 4 and 5 we will outline which outcomes will be considered to be within the same group in terms for the purpose of creating single aggregate indices and carrying out the multiple hypothesis testing adjustments to p-values.

3.9. Creating Aggregate Outcome Measures

As discussed above for each group of outcomes (e.g. child development, quality of the home environment etc.) we will reduce the dimensionality of the outcomes by creating indexed aggregate outcome measures through factor analysis. This has two advantages. Firstly that it creates aggregated measures for the outcomes (intermediate and final) we are most interested in which can aid analysis of the relationships between such outcomes. Secondly, reducing the dimensionality helps the problem of multiple hypothesis testing.

There are several ways to create aggregated outcome measures – simply summing standardised outcome measures, or weighting by the inverse of the covariance matrix (Anderson, 2008). However, we prefer to use factor analysis since it does not make a priori assumptions about the underlying factor structure. The Anderson Index can give higher weight to outcome measures that are less correlated with other outcomes measures (with the justification that due to the low correlation this outcome measure contains more new information). On the other hand using a factor analysis approach gives lower weight (in the primary factor) to an outcomes measure exhibiting low correlation with other measures since this low correlation is likely to indicate this measure is less related to that underlying factor. Since we do not have strong priors on the relative quality of our various measures or the extent to which each is measuring the same underlying construct, or factor, we prefer the factor analysis approach.

We will use STATA's factor command, using the principal factor method. We will fist construct as many factors as there are with eigenvalues larger than 1. We will then examine the factor loadings matrix and discard outcome measures that have loadings of below 0.4 on every factor. We will then reconstruct (if necessary) these factors using this (weakly) smaller set of outcome measures. In cases where this method directs us to keep more than one factor we will keep and estimate the impact of the intervention on both, adjusting for multiple hypothesis testing using the Romano-Wolf method. In these cases we will assess qualitatively if there are clear interpretations that can be given to the different factors that emerge based upon the factor loadings.

3.10. Heterogeneous Effects

We plan to test whether the home visiting intervention had heterogeneous effects on child development outcomes (those listed in Section 4.1) over various dimensions (listed in

Section 4.3). We will test for evidence of heterogeneous effects by using OLS to estimate a linear regression model of the following form where H_{ij} is a dummy variable either equal to 1 or 0 depending on the characteristic we are testing for heterogenous effects over (e.g. 1 for male, 0 for female).

$$y_{ij} = \beta T_j + \rho T_j H_{ij} + \gamma X_{ij} + \delta X_{ij} H_{ij} + \varepsilon_{ij}$$

In this specification our estimated effect of the intervention on the group for which $H_{ij} = 0$ will be β and on the group for which $H_{ij} = 1$ it will be $\beta + \rho$. Therefore, our estimated difference between the effects for the two groups will be ρ and our test of heterogeneous treatment effects for outcome y_{ij} over characteristic H_{ij} will be a test of the null hypothesis $\rho = 0$.

3.11. Binary outcomes with limited variation

For any binary outcome variables we will first look at the variation in the variable. If variation is very limited then there is little power to be gained from looking at this outcome so including it in any aggregate index or group of outcomes for multiple hypothesis testing reduces overall power. In order to limit noise caused by variables with minimal variation, questions for which 95 percent of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests. In the event that omission decisions result in the exclusion of all constituent variables for an indicator, the indicator will be not be calculated.

4. Summary of all Hypotheses to be Tested

We have split the hypotheses we will test when evaluating the impact of the Home Visiting intervention into five conceptual groups. They are summarised in the table below. Each hypothesis is then discussed at length in sections 5 and 6.

PRIMARY OUTCOMES	
A	**Impacts on final outcomes directly targeted by the intervention:** We hypothesise that the Home Visiting intervention will have positive average effects on the following four domains of child development: - A1 – cognition - A2 – receptive language - A3 – expressive language - A4 – fine motor.
HETEROGENEITY IN EFFECT ON PRIMARY OUTCOMES	
B	**Heterogeneity of Impacts:** We hypothesise that effects of the home visiting intervention on children's developmental levels may differ by gender (hypothesis B1). -

SECONDARY OUTCOMES	
C	**Impacts on determinants of child development (intermediate outcomes) directly targeted by the intervention:** We hypothesise that the Home Visiting intervention will have impacts on the following determinants of child development that were directly targeted: - C1 – quality of the home stimulation environment (play activities and play materials) - C2 – maternal time spent on high stimulation activities with children - C3 – mother's knowledge of child development

Table 1: Main hypotheses to be tested

As shown in the Table 1, in our main analysis we distinguish between four main groups of hypotheses: impacts on primary outcomes, heterogeneity in the impacts on primary outcomes and impacts on secondary outcomes. primary and secondary outcomes are outcomes that the intervention directly sought to affect – e.g. the number of play activities that household members performed with the target child or the child's cognitive development. Our primary outcomes are measures of child development directly targeted by the intervention. Our secondary outcomes are measures of determinants of child development that our intervention directly targeted. Our In addition to primary and secondary outcomes our main analysis will also include testing the hypothesis that impacts on primary outcomes are heterogeneous by gender.

In addition to these main hypotheses we also plan to perform test additional hypotheses that we consider to be 'exploratory' and are detailed in Table 2. In contrast to our primary and secondary outcomes, which are outcomes that the intervention directly sought to affect, we also consider impacts on outcomes that were not directly targeted by the intervention but that we might think would, nevertheless, change as a result of it. These indirect impacts might occur because the intervention shifts decision making within the household to put a higher weight on child welfare or because the intervention improved maternal wellbeing. Given we have far weaker priors over the existence of such impacts and we consider them less important for the assessment of the intervention we term them exploratory. Also in this category we include additional hypotheses relating to heterogeneity in impacts of the intervention on primary outcomes over various child and household characteristics. We do not include more heterogeneity hypotheses in our main analysis for concerns over power.

Within both direct and indirect impacts we distinguish final outcomes from intermediate outcomes. When we discuss final outcomes we are referring to indicators of child development, health and wellbeing (as well as maternal wellbeing) that are, in and of themselves important. When we discuss intermediate outcomes we are referring to outcomes relating to the determinants of final outcomes. Clearly, the distinction is not always clear cut.

EXPLORATORY HYPOTHESES	
D	**Impacts on final outcomes NOT directly targeted by the intervention:** We hypothesise that the Home Visiting intervention may have impacts on the following final outcomes not targeted directly by the intervention: - D1 – anthropometrics

	- D2 – gross motor development - D3 – morbidity - D4 – maternal depressive symptoms.
E	**Impacts on determinants of child development (intermediate outcomes) NOT directly targeted by the intervention:** We hypothesise that the Home Visiting intervention may have impacts on the following determinants of child development that were not directly targeted. - E1 – nutrition (dietary diversity) - E2 – schooling plans - E3 – expenditure on books and toys for children - E4 – labour supply
F	**Additional/exploratory heterogeneity analysis:** We hypothesise that the individual characteristics of children and the household will affect the extent to which the home visiting intervention improves children's developmental levels. We hypothesise that impacts may be heterogeneous over the following dimensions: - F1 – age - F2– initial developmental levels - F3 – parity (firstborn child) - F4 – maternal education - F5 – baseline level of stimulation in the home - F6 – wealth of household - F7 – access to safe sanitation

Table 2: Exploratory hypotheses to be tested

5. Main Hypotheses to be tested

In this section we discuss what we hypothesise to be the direct effects of the Home Visiting intervention – effects on outcomes (both intermediate and final) that were directly targeted by the intervention. We divide these hypotheses into three groups: (A) impacts on final outcomes directly targeted by the intervention (in our case, child developmental outcomes), (B) heterogeneity in these impacts on final outcome by child gender and (C) impacts on determinants of child development directly targeted by the intervention. We discuss additional exploratory hypotheses related to indirect impacts of the intervention and additional heterogeneity analysis in section 6.

5.1. Hypotheses Group A: Impacts on final outcomes directly targeted by the intervention (primary outcomes)

Our intervention directly targeted various different domains of child development, with cognitive, language and fine motor development being key. Thus, our main hypotheses of interest for the effect of the intervention on final outcomes (child development) will be assessed by looking for evidence of impacts on the following scales of the Bayley-III:

- Cognition (A1)
- Receptive Language (A2)
- Expressive Language (A3)
- Fine Motor (A4)

We also measure gross motor development through the gross motor scale of Bayley-III. However, since gross motor development was not a major focus of the intervention we do not include this outcome here in our direct impacts on final outcomes (primary outcomes) and instead include it as an indirect impact (secondary outcome) and discuss it in section 6.1. Due to resource and time constraints we have not collected any measure of socio-emotional of Personal-Social development at follow-up and so we do not include this as an outcome.

We hypothesise that the home visiting intervention will have a positive average effect on each of these four domains of child development. We plan to test each domain separately (correcting p-values for multiple hypothesis testing) and create an aggregate index of child development through combining the measures for each domain through factor analysis using the procedure indicated in section 3.9.

For each of the four domains our primary outcomes measure will be internally standardised Bayley-III scores, constructed as outlined in 2.2.3. For robustness and for comparability with other literature we also estimate the impact on each domain using (i) raw Bayley scores controlling for polynomials in age and (ii) externally standardised Bayley-III scores, i.e. composite Bayley scores constructed using the age-specific conversion tables in the manual, which were derived from the normative sample (representative of the US population) (Bayley, 2006).

For each of the four domains we will control for baseline developmental levels using internally standardised ASQ-3 (see next paragraph for details on standardisation) scores from baseline for three domains – Problem Solving, Communication and Fine Motor. We will not control for Gross Motor or Personal-Social ASQ-3 scores since we are not considering Gross Motor development or Personal-Social development to be a primary outcome and is not contained in this set of hypotheses (see above for reasoning). We will control for all three ASQ-3 scores in the estimation of the effect of the intervention on each of the four measures of development since empirical evidence on the technology of skill formation suggesting that skills in one domain are causal in producing skills in other domains in later time periods (Cuhna et al, 2010). In addition, in a world where we expect skills in different developmental domains to be highly correlated and skills to be measured with error (especially in very simple assessments like the ASQ-3) we gain in power from controlling for measures of multiple domains of child development.

Given the ASQ-3 is structured as a series of discrete tests depending on the age of the child (6 months, 8 months, etc.) there are discontinuities in the average scores children get with age. This means the ASQ-3 is not suitable for non-parametric standardisation with respect to age as we outlined above for the Bayley-III. Therefore, we will standardise baseline ASQ-3 scores simply within each age-specific test. We will first remove tester effects by regressing the raw scores for each scale on dummies for each tester and take the residuals. Then for each age-specific scale (e.g. Problem Solving, 8 months) we will calculate the mean and standard deviation of the residual scores and then create a z-score by first subtracting the mean and dividing by the standard deviation. This should give us standardised scores that are comparable across children of different ages.

In addition to testing the hypothesis that the intervention had a positive effect on each domain separately, we will also aggregate the four domains into an overall index (or possibly indices) of child development. This approach has two advantages. Firstly, it will produce one (or more) index of child development with which we can examine the hypothesis that the intervention had a positive average effect on some global measure(s) of child development. Since we know there are strong interlinkages between different domains of development and high correlations between measures supposedly measuring different domains, creating a single (or reduced number) measure(s) can be more informative about the impact on some overall construct of child development. As discussed in section 3.9 using factor analysis to reduce the dimensionality of the outcome measures allows us to be agnostic, a priori, on the nature of the relationships between different domains of child development and the nature of which domains affect the different measures (it could be that conceptually the domains are distinct construct but some measures – Bayley subscales in our case – are causally affected by multiple domains). It uses the covariance structure we observe to estimate the extent to which different measures are related to one or more underlying constructs and predicts the value of these constructs for each child. The second advantage of aggregating outcome measures using factor analysis is that it will reduce the dimensionality of the primary outcomes (most likely to one, but perhaps more, factor(s)). In the case that it reduces the dimensionality to just a single factor of child development this is a solution to the problem of multiple hypothesis testing. In the case when we keep two or more factors then we will use the Romano-Wolf procedure (as outlined in section 3.8) to adjust these hypothesis tests for multiple hypothesis testing.

In each of the hypothesis tests discussed in this section we will use a one-tailed hypothesis test to test the null hypothesis that the intervention does not have a positive effect of children's development as measured by the outcome in question versus the alternative hypothesis that it has a positive effect (i.e. it increases developmental levels). We choose to employ a one tailed test since we have a strong prior that the intervention will not harm the developmental levels of the children.

5.2. Hypotheses Group B: Heterogeneity of Impacts

Hypothesis B1: Heterogeneity over Gender

We will test for heterogeneity in the impacts of our intervention by gender. Differences in parental behaviours towards girl- and boy- children, usually favouring boys and often termed 'son preference', has long been documented in India. Differences in indicators of child health between children of different genders are often attributed to parents differential behaviours in the determinants of health such as diet, vaccination, medical check-ups or breastfeeding. Whilst corresponding reliable analysis of any differences in cognitive, language or fine motor development, and their determinants, has not been documented, this is an interesting area.

Son preference raises the hypothesis of whether the home visiting intervention will have differential effects between children of different genders. This could go one of two ways. Either, the benefits from the additional knowledge and techniques to improve child development provided by the intervention may be focused on boy-children so we would see a greater impacts on the primary outcomes for boys. Conversely, it might be the case that the intervention lessens the concentration of time and recourses towards boys and thus has a greater impact on the primary outcomes for girls.

We use the specification set out in Section 3.8 to test for evidence of heterogeneity by each gender.

5.3. Hypotheses Group C: Impacts on determinants of child development (intermediate outcomes) directly targeted by the intervention

The intervention aimed to affect our primary outcomes of interest through improving the level of psychosocial stimulation that children experienced in and around their homes through working with mothers and primary caregivers. Thus our set of intermediate outcomes (determinants of child development) directly targeted by the intervention are all either measures of the amount of psychosocial stimulation the target child is exposed to in his or her home environment or measures of the knowledge and understanding of mothers and primary caregivers on the process of child development.

Below we provide a description of how we will construct our measure(s) of each. For hypotheses where we have multiple measures we will both report estimates on each measure individually (correcting p-values for multiple hypothesis testing) and create an aggregate measure using factor analysis.

Hypothesis C1: Quality of the Home Stimulation Environment

The home visiting intervention directly targets the levels of psychosocial stimulation in the child's environment. Our key measure(s) for assessing the quality of the home stimulation environment, are based on the Family Care Indicators (FCI) questionnaire developed by UNICEF. As we did in baseline analysis and following Hamadani et al. (2010) we will construct five subscales: sources of play materials, variety of play materials, play activities, household books and household magazines. Hamadani *et al.* show that, in their sample of 801 Bangladeshi children the subscales of the FCI that were most correlated with levels of child development (as measured through the Bayley-III) were the 'play activities' and 'variety of play materials' subscales. This confirms our priors that these indicators would be the best measures of the stimulation experienced by the child. They are also the ones most closely related to the aims of the intervention. Therefore we will only include these two subscales as intermediate (secondary) outcomes. The two subscales will be constructed following Hamadani *et al.* (2010)*:*

1. **Variety of play materials:** the number of different types of play materials the child has played with in the past 30 days. (Maximum score of 7.)

2. **Play activities:** the number of different play activities the child has done with a household member over the age of 15 in the past 3 days. (Maximum score of 7.)

We will also aggregate these two measures into one indexed measure of the quality of the home stimulation environment. We will do this in two ways. Firstly, through factor analysis (equivalent to just summing the two measures since there are only two). Secondly, we will perform IRT (2 parameter model) on the individual items (14 questions) to estimate the value of one underlying latent construct related to home stimulation. We will then standardise this index non-parametrically by age, since these indices tend to increase with age, using the control group as an anchor.

We collect the FCI at two points during the endline data collection:

- Once at the assessment centre (in the same sitting as the Bayley assessment) by the trained by the trained Bayley assessor. In this collection the questions related to play materials had to be done by recall rather than by inspection as interviewers are instructed to do in the home
- Once in the home at the time of the households questionnaire by the household interviewer

These two assessments are relatively close to one another (generally within one month) but ask about distinct periods of time and are administered by different people. This gives us an opportunity to assess the performance of the FCI – if it performs well in measuring some underlying construct we would expect the two measures to have very high correlations – and to construct an outcome measure which makes best use of the available information. However, in this analysis we will simply consider the data collected by the interviewers in the household since they were able to collect the play materials subscale by observation rather than by report.

Hypothesis C2: Maternal Time Spent on High Stimulation Activities with Children

In the household survey we asked all mothers about their time use during the previous working day (Monday to Friday). The aim was to capture how much time mothers spend each day primarily interacting with their child(ren) (i.e. their child being the sole object of their attention rather than just being present) and to capture how much of this time was engaged in play and games with their child(ren). For each category (e.g. cleaning house) we asked the mother to estimate how much time, in minutes, she had spent doing that activity.

As an intermediate outcome (secondary outcome) we will consider the total time (in minutes) categorised as either 'playing with small children in household' or 'reading or telling stories with small children'. In future research we will consider the complexities of this time use data in more detail.

Hypothesis C3: Maternal Knowledge of Child Development

We measure maternal and caregiver knowledge of key principals of child development using an adapted and shortened version of the Knowledge of Infant Development Inventory (MacPhee 1983). This tool attempts to measure knowledge on parental practices, child development processes and infant norms of behaviour. Mothers are read various statements and asked to give their opinion on whether the statement "is true", "is partly true" or "is not true". As in our baseline analysis we will use these answers we construct aggregate scores which measure knowledge under the following domains: (1) praising/paying attention to child, (2) punishing child, (3) school readiness and expectations, (4) importance of maternal interactions and play and, (5) age appropriate expectations.

In testing the hypothesis that the intervention increased maternal knowledge around child development we will test the hypotheses that it had positive average impacts on scores for each domain individually (adjusting for multiple hypothesis testing). We will also estimate the impact on an indexed measure(s) constructed through factor analysis as set out in section 3.9.

6. Hypothesised Indirect Effects of Home Visiting

We are aware of models of intra household optimisation that might predict effects on other investments and therefore other outcomes of interest – we include these here even though our intervention did not directly target these.

6.1. Impacts on final outcomes NOT directly targeted by the intervention

Hypothesis D1: Anthropometrics

Height and weight (and measures of these in relation to age and one another) are important measures of children's medium and long term nutritional status and health. We plan to test the hypothesis that the home visiting intervention shifted investments in children (in terms of nutrition and health seeking behaviours) in a way that affected anthropometric measures. We do not have a strong prior on the direction of such an effect and therefore will perform two tailed tests on all measures.

We will look at the impact of the home visiting impact on the following continuous anthropometric measures: weight for age, weight for height and height for age. More specifically we will create z-scores for each of these three measures following the procedure outlined by the WHO Multicentre Growth Reference Study Group (2006) and report the estimated effect of the intervention on these z-scores. Out of the three we see height for age as the most important indicator of long term nutritional status. However, we recognise that weight for age may respond faster in the set time frame of this intervention and so will be particularly looking for impacts here. We will correct our p-values for multiple hypothesis testing when testing this group of continuous measures of anthropometry. We will also aggregate our continuous anthropometric measures into a single aggregate index of nutritional status as measured by anthropometry using factor analysis as discussed in section 3.9. We will also estimate the effect of the home visiting intervention on this aggregate index.

In addition for looking at impacts on these continuous measures of anthropometry we will also look at the impact on the WHO's binary indicators of nutritional status – underweight (weight for age z-score<-2), wasting (weight for height z-score <-2) and stunting (height for age z-score<-2).

Hypothesis D2: Gross motor development

As discussed in section 4.1 we consider gross motor development, as measured by the gross motor scale of the Bayley-III, as a secondary outcome. We will construct and standardise the gross motor scale in an identical manner to the other scales (outlined in section 4.1). In estimating the impact of the intervention on gross motor development we will control for the baseline score for gross motor on the ASQ-III and core child level controls.

Hypothesis D3: Morbidity

For similar reasoning to that outlined above for long term health status as captured by anthropometric measures we also plan to assess the impact of the intervention on measures of morbidity. In our follow-up questionnaire we measure morbidity as binary indicators of whether the target child has experienced the following symptoms during the specified period, as reported by the mother:

- Diarrhoea in the last 7 days (diarrhoea is passage of loose watery stools more than 3 times a day)
- Fever in the last two weeks
- Fever with shivers in the last two weeks
- Cough in the last two weeks
- Cough with short, rapid breaths or difficulty breathing in the last two weeks
- Vomiting in the last two weeks
- Skin rashes in the last two weeks
- Itching sores on feet and legs in the last two weeks
- Indigestion or stomach pain in the last two weeks
- Unusual tiredness in the last two weeks
- Unusual paleness in the last two weeks

We will aggregate these individual binary indicators into one secondary outcome variable – morbidity. To do this we will use a two parameter IRT model which is well suited for estimating underlying latent factors from a set of binary measures using a conditional MLE approach.

Hypothesis D4: Maternal Depressive Symptoms

We measure maternal depressive symptoms using the CES-D 10 point scale. Mothers were asked six questions on whether they experienced different symptoms of depression over the last seven days. For each question mothers were asked to respond with one of four options: (1) `almost never or never (less than one day)', (2) `a few times (between one and two days', (3) `many times (between three and four times)' or (4) `almost all the time (between five and seven days)'. These answers were scored 0, 1, 2 and 3 respectively and the total depression symptom score was created by summing these scores, so the total score was out of a maximum of 30. The binary cut-off for showing significant symptoms of depression is taken to be 10. However given this has not been validated in our setting we will not use this cut-off and instead only estimate the impact of our intervention on the raw score. Note that this is a short screener and the results should be interpreted as indicating symptoms consistent with depression rather than a diagnosis of clinical depression.

6.2. Impacts on determinants of child development (intermediate outcomes) NOT directly targeted by the intervention

Hypothesis E1: Nutrition (dietary diversity)

As in baseline we collect information on nutrition using simple 24 hour recall period where respondents, usually the child's mother or whoever knew most about the care of the target child, were asked to indicate whether the target children had consumed food from each of a list of categories over the past 24 hours. Dietary diversity, defined as the number of foodgroups consumed in a given period of time, is an important indicator of the quality of a child's diet since diverse diets are more likely to contain sufficient quantities of the wide range of nutrients essential for healthy development. We construct a measure of dietary diversity based upon one proposed by Arimond and Ruel (2004)] which was shown to correlate well with broad measures of nutritional status. We place these above foodgroups into the seven larger foodgroups listed below, scoring each child

as a 1 if they consumed some food in this foodgroup in the past 24 hours and as a 0 if they did not:

1) starchy staples (foods made from grain, roots, or tubers); 2) legumes; 3) dairy (milk other than breast milk, cheese, or yogurt); 4) meat, poultry, fish, or eggs; 5) vitamin A-rich fruits and vegetables (pumpkin; red or yellow yams or squash; carrots or red sweet potatoes; green leafy vegetables; fruits such as mango, papaya, or other local vitamin A-rich fruits); 6) other fruits and vegetables (or fruit juices); and 7) foods made with oil, fat, or butter.

We then construct dietary diversity scores by simply to summing the total number of these seven foodgroups consumed by the child in the past 24 hours. Arimond and Ruel *(2004)* use indicators of whether a child has consumed this food three or more times in the past seven days, however we do not have seven day recall data so we will adopt the method they use for Haiti in their study and use an indicator over the past 24 hours.

Our main measure of quality of nutrition will be this dietary diversity score (secondary outcome).

Hypothesis E2: Schooling plans

We will also look at how the intervention affected households' plans for schooling of the target child, for both secondary and primary school. In this area our two secondary outcomes will be predicted yearly schooling costs (including all fees, uniforms, materials, transport etc) for both primary and secondary school. We will also create an aggregate index using factor analysis.

Hypothesis E3: Expenditure on books and toys for children

We will estimate the impact of the intervention on household expenditure on materials that we would consider providing high levels of stimulation for young children. We place this hypothesis in group E (impacts on determinant of child development NOT directly targeted) rather than C (those directly targeted) since the intervention provided materials for stimulation and encouraged households to make their own and did not explicitly encourage them to buy materials. However, we may well expect that households would change their expenditure on such materials as a result of information provided by the intervention.

We will estimate the impact separately on expenditure in the previous 6 months on books and on toys as well as the total (equivalent to factor analysis with two measures).

Hypothesis E4: Labour supply

We will estimate the impact of the intervention on labour supply of both biological mothers and their husbands, measured in hours per week and weekly earnings (collected in module 2). We will not create an aggregate index of labour supply since conceptually it is difficult to think about what this would be capturing as male and female labour may well be substitutes.

6.3. Hypotheses Group F: Additional/Exploratory heterogeneity of Impacts

Here we list the characteristics over which we will test for heterogeneity of effects of the home visiting intervention on child development outcomes (those listed in Section 4.1), as exploratory analysis. Below we describe how we convert each characteristic into a binary indicator. We then use the specification set out in Section 3.8 to test for evidence of heterogeneity by each characteristic.

Hypothesis F1: Age

We are agnostic to whether the intervention will have a larger, smaller or similar impact on child developmental outcomes for children who are older when they begin receiving the intervention (and thus receive it until an older age). However, psychometric tests, such as the Bayley, are often found to be more sensitive for slightly older children and therefore we might expect to more precisely estimate impacts for older children even if the underlying impact on child development is, in fact, the same across the groups.

We will divide our sample into halves by age at baseline to define our cut-off of which children are classed as 'older' and which as 'younger'.

Hypothesis F2: Initial Developmental Levels

We do not have a strong expectation as to the direction of this heterogeneity. On the one hand children with lower initial developmental levels may find it harder to engage with some of the activities (although the levels should be adjusted to the developmental level of the child). On the other hand, children with lower initial developmental levels have more scope for improvement over the period.

Our baseline measure of developmental level is the third version of the Ages and Stages Questionnaire (ASQ-III). We will take standardised ASQ-III scores for the three domains that correspond to our developmental outcomes of interest at follow-up – problem solving, communication and fine motor – and use them to construct an aggregate index of development at baseline using factor analysis, using the method discussed in section 3.9 and taking only the first factor. We will then split the sample in two halves by this index of baseline developmental level and define the lower half as 'lower initial developmental level' and the higher as 'higher initial developmental level'. As further exploratory analysis we will also look at heterogenous effects by the four quartiles of the distribution of initial developmental levels.

Hypothesis F3: Parity

The direction of this hypothesized heterogeneity effect is unclear. On the one hand first born children may benefit more from the intervention is their mothers/ primary caregivers have more time to spend on the stimulating activities introduced in the intervention. However, given later children tend to have lower developmental levels than first borns of the same age, we may think that there is more potential for improvement here.

Hypothesis F4: Maternal Education

We have no strong prior on the direction of this heterogeneity. On the one hand mothers with lower education levels may, for example, find it harder to fully grasp some of the concepts behind the intervention and utilize these in every day practice. On the other hand there may be more scope for improvement in parenting practices with this group.

We measure maternal education by years of education as recorded at baseline. We will analyse heterogeneity over (1) middle school or less (8^{th} standard or less) which corresponds to 52% of the sample and (2) primary school or less (5^{th} standard or less) which corresponds to 30% of the sample.

Hypothesis F5: Baseline Level of Stimulation

We have not strong prior on the direction of this heterogeneity. On the one hand children who experience higher levels of stimulation at baseline may see a lesser increase in the amount of stimulation as a result of the intervention and their developmental level may, therefore, increase less. However, parents who are already providing a higher level of stimulation may be more receptive to the concepts of the intervention and may, therefore, increase the amount of stimulation they provide more.

We will measure baseline level of stimulation using an aggregate factor index of the baseline FCI score (constructed as described in section 4.2.1), standardised by age. We will then divide the sample into two halves based on this baseline score – the lower baseline stimulation half and the higher baseline stimulation half.

Hypothesis F6: Wealth

We are agnostic on the direction of this heterogeneity. On the one hand poorer households may benefit more from the intervention since the concepts and practices encouraged do not explicitly require monetary investments and their children may be starting from a lower baseline and therefore more improvement may be possible. On the other hand richer households may be able to reallocate more resources (both time and money) into child development as a result of the intervention.

We will construct a baseline wealth measure using an aggregate factor index of household asset ownership. We will include all assets we collected information about (bicycle, motorbike, car, fridge, fan, washing machine, cooker, sewing machine, table, chair, cot, mat, tv, mobile phone, laptop and radio). We will keep the largest factor (only one factor has an eigenvalue >1).

Hypothesis F7: Access to Safe Sanitation

We will test whether there is heterogeneity in the effects of the intervention over access to safe sanitation at baseline. We will define safe sanitation as the dwelling being equipped with one of 'toilet connected to sewage system', 'toilet connected to septic tank', 'toilet connected to the drain' or 'pit latrine'. We will also create aggregate village indices of safe sanitation practices and look for heterogeneity over these because of the presence of strong externalities in the effects of sanitation.

References

Andrew, A, Attanasio, O. P., Augsburg, B., Grantham-McGregor, S. M., Meghir, C., Pahwa, S. & Rubio-Codina, M. (2014). Early Childhood Development in the Slums of Cuttack , Odisha, India: Baseline report.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American statistical Association*, *103*(484).

Arimond, M., & Ruel, M. T. (2004). Dietary diversity is associated with child nutritional status: evidence from 11 demographic and health surveys. *The Journal of nutrition*, *134*(10), 2579-2585.

Attanasio, O. P., Fernández, C., Fitzsimons, E. O., 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.

Bayley, N. (2006). Manual for the Bayley scales of infant and toddler development, third edition, Pearson Education , Inc..

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883-931.

MacPhee, D. (1983). The Nature of Parents' Experiences with and Knowledge about Infant Development.

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*(1), 23.

Romano JP, Wolf M. 2005. Stepwise Multiple Testing as Formalized Data Snooping. Econometrica; 73(4):1237-1282.

Romano JP, Wolf M. 2005. Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. Journal of the American Statistical Association; 100(469):94-108.

WHO Multicentre Growth Reference Study Group., WHO Child Growth Standards based on length/height, weight and age., Acta paediatrica. Supplementum 450 (2006) 76_85. doi:10.1080/08035320500495548.