Methodological challenges associated with assessment of the impact of a parental education program on child development and psychosocial stimulation: Findings from a secondary analysis of data from a quasi-experimental program evaluation in Tijuana, México

Beth Lee Mynar

A thesis

submitted in partial fulfillment of the

requirements for degree of

Master of Public Health

University of Washington

2015

Committee:

Paula E. Brentlinger

Annette E. Ghee

Program Authorized to Offer Degree:

Department of Global Health

University of Washington

Abstract

Methodological challenges associated with assessment of the impact of a parental education program on child development and psychosocial stimulation: Findings from a secondary analysis of data from a quasi-experimental program evaluation in Tijuana, México

Beth Lee Mynar

Chair of Supervisory Committee:
Paula E. Brentlinger, Clinical Assistant Professor
Department of Global Health

Background: The director-general of WHO recently declared that attainment of a child's full developmental capacity is a human right, and is critical to "equitable prosperity and sustainable progress of societies." Inadequate psychosocial stimulation has emerged as an important modifiable risk factor for poor child development, but best practices for scale-up of practical and cost-effective interventions are still undefined. We aimed to determine whether a parental education program developed and administered by World Vision (WV) under field conditions resulted in improved child development.

Methods: We conducted a secondary analysis of de-identified data derived from a quasi-experimental study conducted in two communities (intervention and comparison) in Tijuana, México. The intervention consisted of a parental education program. Data on child milestone attainment and household characteristics were obtained from maternal self-report in baseline and post-intervention household surveys administered to randomly selected households with young children. The primary analysis employed a difference-of-

differences approach to compare change in the proportion of children meeting milestones between baseline and endline assessments in intervention vs. comparison communities, in multivariate logistic regression models controlling for potential confounders.

Findings: Data were available for 396 households with children 6 to 23.99 months of age (comparison: 120 baseline, 36 endline; intervention: 109 baseline, 131 endline); 375 were retained in the final multivariate analysis. In the final model, the odds ratio (OR) for attaining all developmental milestones evaluated at endline vs. baseline in intervention vs. comparison communities was 3.70 (p = 0.02, 95% confidence interval [CI] 1.27, 10.78) in a model including child age, maternal depression, and provision of iron-rich foods (all significantly associated with milestone attainment). The final model was derived from unplanned analyses, because observed odds ratios for associations between covariates within the psychosocial stimulation domain and milestone attainment varied significantly and inexplicably by arm and time in the prespecified multivariate model. There was no statistically significant association between household-level exposure to the intervention and milestone attainment. Examination of the comparison community at endline revealed that it appeared to differ in significant and fundamental ways from all other study subgroups for reasons we cannot fully explain within the constraints of this data set.

Interpretation: Assessment of the impact of the WV program was significantly limited by methodological constraints related to study design and implementation. While findings of multivariate models suggest that the WV intervention was associated with improvements in child development within the intervention community relative to comparison community, we are unable to definitely attribute these to the WV

intervention. Ultimately, the overall relationships regarding psychosocial stimulation, exposure to the intervention, and child development within this data set did not form a consistent pattern. While the methodological weaknesses of this study prohibit us from drawing robust inferential conclusions, this analysis provides important insights into some of the common challenges faced by research seeking to evaluate early child development programs. We end with a section of recommendations for future studies that draw attention to some of the key lessons learned in this study in an effort to help future studies avoid similar pitfalls.

Background

It has been estimated that over 200 million children under the age of five are not meeting their development potential in low- and middle-income countries, largely as a result of health and social factors related to poverty.[1] This has long-term consequences not only for these children, but also for the larger community, as it is increasingly being recognized that developmental disturbances in early life can have negative repercussions on cognitive, behavioral, and economic attainment even into adulthood.[2–4] Through this mechanism, early developmental disturbances have the potential to increase already pronounced economic disparities between low- and high-income countries.[5] The importance of early child development was highlighted by the World Health Commission on Social Determinants of Health treatise, published in 2008, which concluded that to reduce health inequalities worldwide, one of the highest priorities must be to provide all children with the best start in life.[6] The director-general of WHO recently declared that attainment of a child's full developmental capacity is a human right, and is critical to "equitable prosperity and sustainable progress of societies."[7]

Given the increasing recognition of the importance of enabling children to attain their developmental potential, considerable attention and funding has been directed toward research on early child development (ECD) programs. Child development is conventionally assessed in four developmental domains (gross motor, fine motor, language, and social-emotional), using developmental milestones to provide standards indicating the age at which a child can be expected to attain a certain skill.[8] Child development research has shown that the first two years of life are perhaps the most critical period in the developmental process.[9–11] During this window, social, nutritional and biological factors heavily influence development, and set the trajectory for an individual's health vulnerability throughout life.[12] The importance of this particular window is grounded in brain development processes, and relates to the neuronal pruning of unused circuits that occurs after age two.[9] The interplay of the numerous social and biological factors during child development is highly complex, and considerable energy is currently being invested in studies that seek to better understand how these factors impact a child's development in each of the four developmental domains.[10]

As preparation for of this thesis project, over 60 papers investigating risk and protective factors for child development, and/or interventions that might promote child development, were retrieved and reviewed. This limited literature review found that commonly accepted risk factors for poor child development include, but are not limited to, nutritional factors (stunting, intrauterine growth restriction, iodine deficiency, and iron-deficiency anemia), illness (infectious and non-infectious), societal factors (institutionalization and violence), maternal depression, and inadequate psychosocial stimulation.[2,12] These factors have been shown to be associated with poor child development across cultures, suggesting an important role in the development process. Several protective factors have also been identified, including maternal education and breastfeeding.[2]

Inadequate psychosocial stimulation has only recently emerged as an important modifiable risk factor for child development. Several studies have demonstrated improvements in child development through a variety of different interventions focused on increasing psychosocial stimulation during the critical period.[13] These studies took place in low- and middle income countries, and included community based education

programs held in centralized locations,[14,15] interventions that took place in a hospital or clinic context,[16,17] home-based one-on-one teaching sessions focused on developing parenting skills,[18–22] and a combination of these strategies. Systematic reviews comparing studies suggest that effect sizes may be greater for interventions that include both parents and children,[5] and that combine group-based sessions with home visits.[23] However, psychosocial interventions of many forms have shown beneficial effects on child development.

Based on the literature, the category of child development most amenable to improvement through increased stimulation is cognitive development (including both language and social-emotional development). Studies conducted in both individual and group settings, and across cultures, have shown clear benefits to cognitive development,[17,21,22,24] and to a lesser extent motor development.[16,18] The majority of child development research looks at changes in developmental outcomes over short time frames, however the few longitudinal studies that exist have shown long-term cognitive, behavioral, and economic benefits.[3,25,26] A 22 year follow-up study of a four-arm randomized control trial in Jamaica in 1991 looked at the impact of psychosocial stimulation and food supplementation on child development and found that during adulthood, children who had been in the stimulation arms had a higher IQ, greater educational level, reduced participation in violent crimes, and greater economic attainment (in the form of wage earnings) in comparison to individuals in the no-stimulation arms.[3,27]

In addition to providing ample evidence for the importance of psychosocial stimulation to child development, the available literature highlights several significant limitations in the field of child development research. Specifically, research investigating child development and psychosocial stimulation is hampered by inconsistency in regard to the definition and measurement of these factors.[28] As will be discussed in greater detail below, this both limits cross-study comparisons and interpretation of findings, and also makes it difficult to reproduce findings in future studies or rollout interventions based on study conclusions.

The fact that there are numerous techniques used to measure child development, many of which are both time-consuming and costly, is another notable challenge for child development research. The majority of the studies reviewed assessed child development through either maternal self-report (in the form of a questionnaire or interview), or direct infant observation by a highly trained healthcare worker (who typically evaluated the infant according to a developmental scale). However, the questionnaire or scale used to assess the child's developmental status often varied by study. Two commonly used child development scales include the Bayley Scales[29] (with subscales for cognitive, language, fine motor, and gross motor development) and the Griffiths Mental Development Scales,[30] each of which has been translated into numerous languages. While use of either of these scales ensures that child development is measured against a set of core standards, these scales have been normalized for American or European populations and are often not validated in the relevant study populations.[23] Furthermore, these measurement tools require a trained observer to evaluate the child, and therefore are costly and/or time-intensive to administer. Adaptation of scales to specific cultural contexts may also be expensive and labor-intensive. This can be a significant barrier for work in developing countries, where insufficient funding and limited resources can be prohibitive. Research seeking to develop child development tools that are appropriate for resource limited

settings is ongoing at this time. For example, the Kilifi developmental scale has now been adapted for use in several settings in sub-Saharan Africa,[31–33] but as of yet there is no commonly accepted standard.[31,32] What is lacking in the field of child development research is a validated concise assessment tool that is based on caregiver report, or brief evaluation by field workers, and is applicable or can easily be adapted across cultures.

Similarly, there is no universally agreed upon mechanism for assessing the content or quality of psychosocial stimulation received by a child.[23,28,34] Techniques used in the literature include, but are not limited to, measuring the total amount of time parents spend interacting with their infants, or investigating whether increased levels of specific activities (such as listening to music, reading books, or playing with toys) are associated with improved development.[13] In addition, scales have been developed seeking to measure the overall level of psychosocial stimulation received by a child, such as the Home Observation for Measurement of the Environment (HOME).[35] The version of the HOME scale for children under 24 months includes 45 items that are assessed through observation of the mother and infant, notation of play materials in the home, and a battery of questions regarding the child's exposure to places, people and conversations.[23] A briefer version, called the Family Care Indicators scale, has also been created and validated in Africa and South Asia.[23] However, as with child development, these scales are often not re-validated for use in different study populations, thus drawing into question their reliability as an evaluation tool across cultures. Additionally, the high cost and time demands associated with administration of these scales can be prohibitive.

In addition to revealing the limitations specifically associated with the measurement of child development and psychosocial stimulation, the literature review highlighted the dearth of information regarding integration of multiple child development interventions outside of a research setting. While many studies have looked at the independent impact of key interventions (such as nutritional supplementation or psychosocial stimulation) on child development, a smaller number of studies look at the effect of combined nutrition and stimulation interventions.[36,37] Even fewer studies look at the impact of integrated early child development programs that seek to mitigate a wide array of risk factors, beyond simply poor nutrition and inadequate psychosocial stimulation.[38] Therefore while child development research has enabled us to describe different elements that contribute to successful ECD programs, there is limited information regarding how these elements should be combined in an integrated intervention or the impact of ECD programs outside of a research context.[28]

The present work sought to address some of these questions by assessing the impact of an integrated ECD program that taught parenting skills in a community in México. The program was run by World Vision (WV) México, the national office of the global Christian development and relief organization, and consisted of a series of parental education group sessions emphasizing topics of known importance during the critical developmental window. The goal of the WV intervention was to show improvements in child developmental outcomes through a community based integrated ECD program. Given the limitations associated with measuring child development and psychosocial stimulation, WV investigators chose to develop their own assessment tool to evaluate the ECD program's impact. The tool consisted of an extensive questionnaire (see methods section) assessing markers of child development in addition to maternal, household, and child factors of known importance to child development.

This thesis consists of a secondary data analysis that emerged from the WV program evaluation and builds on prior analytic work, both by WV and by a previous MPH student (Mukerjee K, 2013).[39] The original analytic work performed by WV[40] did not adequately enable WV to evaluate its intervention for two key reasons: a. the measure of child development used during the initial WV analysis did not adequately adjust for the association between child age and milestone attainment, thus does not accurately reflect the child's developmental status; and, b. comparison between intervention and control groups was not done via a multivariate difference-of-differences analysis, and thus did not fairly reflect changes in development status attributable to the WV intervention after adjustment for confounders and effect modifiers. The work conducted by Murkerjee[39] evaluated baseline data only, and thus was unable estimate the impact of the WV program.

The initial goal of this thesis work was two-pronged: a. to answer the primary study question, did the WV intervention result in improved developmental outcomes, based on the appropriate difference-of-differences analysis; and, b. to better understand the relationship between psychosocial stimulation and child development within this data set. Secondarily, this thesis work also sought to shed light on the strengths and weaknesses of the evaluation tool developed by WV. However, as will be revealed over the course of this thesis document, the lessons we will truly take away from this thesis work are largely methodological in nature, and answers to inferential questions remain elusive.

Methods

Parent study design

This thesis consists of a secondary data analysis of previously collected data gathered as part of a WV parent study investigating the impact of a WV intervention on child development outcomes. The quasi-experimental parent study employed a repeat cross-sectional design, based on two waves of community-based data collection (baseline and endline) in one intervention and one comparison (control) community. A quasi-experimental design was selected (over a randomized or cluster-randomized controlled trial) given that the intervention took place at the level of the community and financial/logistic constraints limited the size of the overall study.[41–43] Randomization of individual households to intervention vs. no intervention within one community was not feasible given the possibility of intra-community contamination (in the form of information sharing between neighbors). Importantly, the two communities included in this study were not randomly assigned to comparison or intervention groups.

Parent study setting

Both intervention and comparison communities were located in El Cañón del Sáenz, in the municipality of Tijuana, México. Sampling for the parent study occurred at two time points – pre- and post-intervention. The baseline assessment took place from June to August of 2010, and the endline took place between September and December of 2012. The specific study sites were selected because they were the two locations in Tijuana where World Vision had established a programming presence at the time. There

was no formal assessment of sociodemographic similarity between the two communities prior to site selection.

Participants

At both time points, eligible households were defined as households with pregnant women and/or children less than three years of age. Inclusion criteria included: a. mothers between the ages of 18 to 49 years of age; b. Spanish-speaking interviewees; and c. absence of a disability rendering interviewees unable to answer questions with veracity or participate in the informed consent process (e.g. deaf without sign language interpreter, mute, severe developmental or mental health disability).

In both baseline and endline samples, participants were selected via simple random sampling of households following census activity. Consistent with the repeat cross-sectional design, households that had participated in the baseline survey were neither intentionally included nor excluded in the endline survey. Similarly, households that had participated in the intervention were not specifically included nor excluded in the endline survey.

In the comparison community at endline, interviewers had difficulty recruiting an adequate number of participants, thus modest financial incentives were offered to encourage study participation. Incentives were not originally contemplated in study design, and no incentives were offered in the comparison community at baseline, or in the intervention community at either time point.

Intervention

The WV program consisted of a series of hour-long parental education sessions. These group sessions were held in a variety of community-based venues and collectively covered numerous topics important to child development, including but not limited to child nutrition, play and other forms of psychosocial stimulation, child development milestones, personal hygiene practices, and the importance of positive discipline to promote development. The WV program was divided into chapters, each chapter covering the above topics as they related to a specific age group. The intention was to form caregiver cohorts (based around child age), with each cohort participating in monthly or biweekly sessions to cover relevant chapters. The first chapter consisted of 25 sessions for infants between birth and nine months, the second chapter consisted of 26 sessions for children between ten and 36 months, and the third chapter consisted of 21 sessions targeting children up to five years of age. The program was open to all households in the intervention community, and attendance by parents and/or other caregivers was voluntary. The program did not include household visits. After the endline survey, the program was offered in the comparison community.

The parental education program was originally intended to follow a specific curriculum designed by WV. This curriculum was based on issues of known importance to child development, and outlined specific goals for each session. However, a qualitative evaluation of the program documented that though the session leaders followed the general structure of the overall program, they did not follow the sequence of topics as defined in the curriculum.[44] Instead, the topics covered at each meeting were dictated by group demand. The sequence and range of topics actually discussed was not recorded and is not known.

The WV program was rolled out over a one-year period beginning immediately after the baseline assessment. Hence the program deployment period is thought to have extended over approximately 18 months, a relatively brief interval.

Data source

As no standardized instrument existed for the purpose of evaluating interventions of this nature at the community level, WV developed the early childhood care and development (ECCD) program questionnaire to evaluate its intervention through investigation of changes in child development and parenting practices in each community between the two time points. The ECCD questionnaire included questions assessing child development and infant stimulation, as well as multiple other health-related domains, including maternal demographics, maternal depression, household composition, child nutrition and feeding practices, and household sanitation. The questionnaire was developed for this study specifically, and was informed by World Health Organization (WHO) multi-country evaluation of motor milestone attainment[45] and other published child development literature. It was not previously validated, with the exception of the Center for Epidemiologic Studies Depression Scale[46] (20 questions assessing symptoms of depression on a 60 point scale), which had been validated within Spanish-speaking populations in the United States but not in México.[47]

The same questionnaire was used in both the baseline and endline surveys, except for an additional battery of questions designed to measure exposure to the WV education intervention at endline. Medical students from the Universidad Autónoma de Baja California were trained to administer informed consent and conduct face-to-face interviews in Spanish with mothers of index children. The same team of interviewers and field manager conducted interviews in both communities, but different teams were recruited for baseline and endline field work. All data were gathered through maternal self-report, and the interviewers in this study did not independently observe child behavior, measure height or weight of infants, or collect blood samples from participants.

Human subjects

The original World Vision study was approved by the Human Subjects Review committee at Universidad Autónoma de Baja California. The secondary data analysis (this thesis work) used only de-identified data and was determined to be exempt from review by the University of Washington Human Subjects Division (Fall, 2013).

Variables for secondary data analysis: outcome(s), exposure(s) and other covariates

As discussed in the introduction, there is no commonly accepted brief survey method to assess or compare child development status or exposure to psychosocial stimulation across groups or time points, therefore WV was required to develop its own method for evaluating these domains within its survey tool. Given stringent financial and logistic limitations, the WV team was limited to only a few questions specific to child development (the primary outcome of interest) and psychosocial stimulation (the primary exposure of interest). One of the key first steps in this thesis project's exploratory analysis was to define variables that described psychosocial stimulation and child development status.

The ECCD questionnaire questions assessing different elements of a child's developmental status fell into two separate, yet overlapping, categories: overall language acquisition, and achievement of selected developmental milestones. Mukerjee's work[39] investigating associations in the WV baseline data developed a variable summarizing a child's overall development status that was based on the attainment of 12 milestones (not including vocabulary, as the continuous vocabulary measure did not conform to the dichotomous milestone measures). This variable was termed "milestone attainment," and was represented by a dichotomous variable that denoted "attained all milestones evaluated in study questionnaire" vs. "failed one or more milestones evaluated." The specific milestones assessed for the 6 – 11.99 age range included holds up head independently, rolls over independently, sits independently, smiles at a family member, follows sound or music, crawls, stands, responds to their name, and makes or reproduces sounds. The 12 – 23.99 age range assessed three additional milestones: follows easy instructions, recognizes and names sounds, and walks confidently. As discussed by Mukerjee,[39] the milestones assessed in the ECCD questionnaire did not always align with commonly accepted child development milestones; for example, "recognize and names sounds" and "follows simple instructions" do not correspond with a designated WHO milestone.[39] See work by Mukerjee[39] for a discussion of the strategies used to create the "milestone attainment" variable as well as the limitations resulting from the occasional lack of correlation with WHO milestones, over representation of motor milestones at the expense of other developmental domains, and poor matching between milestones selected and the age range of children evaluated. The other primary outcome measure, language acquisition, was not addressed in the earlier thesis project, and was measured in the ECCD questionnaire by a single question asking the number of words spoken by the child.

The domain of psychosocial stimulation was measured in the ECCD questionnaire through description of the diversity of stimulation activities taking place in the home and the total amount of time a child received stimulation. To ascertain stimulation type, the ECCD questionnaire asked respondents what types of stimulation activities they, or another family member, engaged in with the child: play with or without using objects, sing, listen to music, speak, read, or massage/body movements. Importantly, the questionnaire did not address the quality or quantity of each activity type, nor did it specify the time period over which specific activities occurred. To assess stimulation time, the ECCD questionnaire asked how much time the respondent, or another family member, spent in psychosocial stimulation activities with the child during the prior day, with a total of five possible responses: no stimulation, less than half an hour, half-hour to an hour, one to two hours, or greater than two hours. The questionnaire did not seek to describe the form(s) of stimulation that occurred during the specified time period.

The ECCD questionnaire also described numerous child, maternal, and household characteristics of known importance to child development. Each of these variables was previously defined by Mukerjee[39] and included sex of infant, dietary diversity, diet containing iron-rich foods, maternal depression (as measured by self-administered Center for Epidemiologic Studies Depression Scale), maternal employment outside the home, maternal employment as a maquiladora, maternal age, maternal education, and single parent household. These variables were defined based on WV measurement standards, for which the indicators were benchmarked to UNICEF Multiple Indicator Cluster

Survey (MICS) (see prior thesis work).[48] Table 1 summarizes the format and definition of key variables incorporated into our analyses.

At endline, participant exposure to the intervention was also assessed. Specifically, the questionnaire asked if participants had heard of an education program run by WV teaching parents about the care of children, if they knew the name of the program, if they had participated in the program, and if so, how many sessions had the participants attended. If applicable, the questionnaire went on to ask what topics the participants recalled covering during the WV sessions.

Sample size

Sample size and power calculations were completed prior to the initiation of the parent study and were based on the following factors. It was assumed that the intervention would result in a change of 15 percentage points between baseline and endline in the outcome indicator in question within the intervention community, while no change would occur in the comparison community. Since plausible baseline values were unknown for key indicators in this study, sample size calculations were conducted at a range of potential baseline values. The formula assumed 80% power and an alpha of 0.05, and an additional 10% was added to the calculated sample size to account for participants who refused to participate or could not complete the survey for other reasons. The outcome indicator for the power calculations was not specified. Sample size calculations did not consider the effect of the cluster design on sample size requirements, or the potential clustering effect.

Calculations determined that a minimum sample size of 300 was recommended in each study arm at each time point (baseline comparison, endline comparison, baseline intervention, and endline intervention) in order to detect a change in indicators with an assumed baseline value of approximately 25%. However, this sample size was not obtained.

Statistical methods

This secondary data analysis was designed *a priori* to be conducted in a series of sequential steps initially seeking to define key variables and ultimately culminating in the creation of a multivariate model characterizing the relationship between child development, psychosocial stimulation, and other important covariates within the parent dataset. Each analytic phase is described in detail later in this section. The early phases sought to determine the most appropriate definitions for variables falling into the child development or psychosocial stimulation domains given the constraints of small sample size and ambiguous wording of ECCD questionnaire (as noted above). Later phases focused on identifying potential confounding variables to be considered at the subsequent multivariate analytic phase. The final multivariate model employed a difference-of-differences approach to answer the primary study question: "did the WV intervention lead to a greater measurable change in child development in the intervention group as compared to the comparison group over time?" and the secondary question was "if so was the causal pathway through psychosocial stimulation (that was associated with exposure to the WV intervention)?"

Although the parent study included households with children from 0 to 36 months of age, this secondary data analysis only included a subset of the original sample(s). This

analysis was restricted to children between the ages of six and 23.99 months (if the outcome was milestone attainment), or 12 and 35.99 months (if the outcome was language acquisition). Analysis of milestone attainment excluded children under six months of age due to the small number of children with available data in this age group, and due to the weak discriminating power of the survey instrument to assess the rapid developmental gains in this age range. We also excluded children over 23.99 months of age because the ECCD questionnaire included insufficient questions assessing milestone markers for this age period. Analysis of language attainment only included children between the ages of 12 and 35.99 months as word count was only assessed in this age range.

Because child age is the principal driver of all child development, every modified bivariate and multivariate model used in this secondary data analysis included a covariate that adjusted for child age, as this adjustment was considered fundamental to describing the relationship between all variables and child development. All models also included a covariate for "questionnaire" that adjusted for the fact that the 6 – 11.99 month questionnaire did not assess all milestones assessed in the 12 – 23.99 month questionnaire. Additionally, where appropriate, all models included an interaction term for time and study arm. Questionnaire and the time/arm interaction term were the core design variables in the bivariate and multivariate analyses.

All analyses were conducted in Stata (version 13) [Stata Corporation, College Station, Texas, USA]. Analyses were conducted in pre-specified phases, to be described below.

Phase 1: defining candidate variables to describe child development status, the primary developmental outcome of interest

The preliminary analysis began with an investigation of whether milestone attainment or language acquisition was a more appropriate primary outcome within the child development domain within this data set. Although it was initially hoped that language would prove a fruitful surrogate for cognitive development, initial investigations revealed that the variable had a highly positively skewed distribution (the range was 0 to 200, but half [216 (50.9%)] of index children were reported to know 10 or fewer words) and many missing values (82 missing values out of a total of 444, or 18.5%). This distribution meant that the small number of outliers at the upper end carried too much analytical weight, and the vast majority of observations at the lowest end were so close together that developmentally distinct categories could not be created while preserving any potential differences between arms of the study. This prevented language acquisition from functioning as a useful development marker, and therefore, we selected dichotomously measured milestone attainment as the primary development outcome of interest for this study.

Regarding milestone attainment, Mukerjee's analyses determined that the most appropriate measure of milestone attainment was a dichotomous variable denoting passed all milestones assessed vs. failed one or more milestones.[39] There are two primary reasons for this, resulting from factors inherent to child development as well as limitations specific to the WV study design. First, the development window during which a child can be expected to attain a certain milestone is often long, resulting in broad confidence intervals around the expected child age for attainment of each milestone. As a

result, the attainment of a group of milestones is more informative than the attainment of individual milestones. Additionally, the ECCD questionnaire assessed too few milestones and too few developmental domains to permit identification of comparable milestone subsets for each age group of interest (e.g. subgroups subdivided by age in months), thereby necessitating grouping all ages. For further details regarding the development of the dichotomous measure of milestone attainment please see work by Mukerjee.[39]

Phase 2: description of study populations

We used percentages and medians to describe (see Table 2) important characteristics of mothers, infants, and households, stratified by time, arm and exposure to the WV program. As milestone attainment had been selected as the primary developmental outcome, the observations included in Table 2 were restricted to households with children 6 – 23.99 months, who were not deemed ineligible (defined as missing age or milestone information).

Baseline comparisons of main variables were previously done by Mukerjee[39] and were not repeated in this secondary data analysis. However, the present analysis did compare baseline characteristics of the two groups (using χ^2 and T-testing) for important variables not previously evaluated (primarily in the psychosocial stimulation domain).

Phase 3: evaluating exposure to the intervention

Exposure to the intervention was assessed in the ECCD questionnaire at a variety of different levels, ranging from a general awareness of the existence of a WV program within the community (without actually having attended the program) to an ability to recall specific topics covered in sessions attended by participants. In the original WV analysis plan, the goal was to use exposure to the intervention to help shed light on the direct and indirect effects of the WV program. The plan was to compare intervention-exposed, intervention-unexposed, and comparison communities to assess for possible secondary effects of the intervention that were mediated through the community, such as potential benefits that resulted from exposure to someone who had participated in the WV intervention.

Unfortunately, the usefulness of exposure to the intervention as a variable was severely limited by the small number of study participants reporting even small degrees of exposure to the WV intervention. The lowest level of exposure explored - having heard of the existence of a WV program but without having attending any sessions – seemed unlikely to provide useful information regarding program impacts given the almost negligible level of exposure to the intervention. For this reason, simple awareness of the program was discarded as a potential definition of exposure status. The second lowest degree of exposure to the intervention assessed in the ECCD questionnaire was having participated in one or more sessions. Only 27 subjects responded affirmatively to this question, and even fewer recalled specific session topics covered (the highest level of exposure measured). Given the small number of individuals who were exposed at the lowest meaningful level measured, having participated in one or more sessions was the only viable definition for exposure to the WV intervention in this analysis. However, when this variable was tested in modified bivariate logistic regression models (controlling for child age and design variables) looking for associations between the WV intervention and milestone attainment, there was no significant association noted between

exposure status and developmental outcomes. For this reason, direct (individual or household) exposure to the intervention was not included in any multivariate model, and only appears in the descriptive analysis in Table 2.

Phase 4: defining candidate variables within the psychosocial stimulation domain, the primary exposure(s) of interest

In regard to defining the primary exposure of interest (psychosocial stimulation), we evaluated several alternative candidate variables, as possible measures of the psychosocial stimulation domain. We created different forms of stimulation activity and stimulation time variables and tested each in logistic regression models (controlling for child age and design variables) using milestone attainment as the dichotomous primary outcome. A marginally statistically significant association between candidate stimulation variables and milestone attainment was defined as a p value <0.1. This approach allowed us to consider a broader range of covariates at the subsequent multivariate modeling stage. We created stimulation time variables in both a categorical version and a linear approximation version (see Table 1), and the version with the strongest association with milestone attainment was a categorical version containing three ranked categories: no stimulation, up to 2 hours of stimulation, and greater than 2 hours of stimulation. Similarly, we created and compared multiple candidate stimulation activity variables, including a "diversity of activities" measure (a count variable that was derived by adding up the number of stimulation activities reported) as well as dichotomous variables representing participation vs. no participation in each of the six stimulation activities. Though the diversity of stimulation activities variable was marginally associated with milestone attainment, analysis of each of the stimulation activities independently revealed that bivariate associations with milestone attainment were only statistically significant for play and (marginally) for listening to music (see Table 3). No other stimulation activity demonstrated a statistically significant relationship with milestone attainment. Based on these findings, we determined that the most appropriate variables within the psychosocial stimulation domain included stimulation time (categorical version), play (dichotomous) and listening to music (dichotomous).

Phase 5: combining variables from the child development and psychosocial stimulation domains into one core model

Phase five was the first step toward creating what would ultimately become the multivariate model. We undertook a series of exploratory logistic regression analyses looking at the associations between milestone attainment (the dichotomous outcome variable) and play, listening to music, and/or greater than 2 hours of stimulation time (candidate variables in the psychosocial stimulation domain defined in phase 4). During this phase, we included all four study groups (defined by arm [intervention/comparison] and time [baseline/endline]) in one model that looked for overall associations within the data set. Each model was adjusted for child age, questionnaire, time and arm.

We added each stimulation covariate in a stepwise fashion to a model containing milestone attainment as the primary outcome. All possible combinations of play, listening to music, and stimulation time were tested, and a change in the odds ratio of greater than 10% from the baseline association (between each individual stimulation covariate and

milestone attainment) was used to suggest the presence of a complicating factor, such as confounding or collinearity.

Based on these analyses, a model containing play and listening to music emerged as the seemingly most appropriate core model as each stimulation activity maintained its independent association with milestone attainment in this model. The addition of stimulation time to a model containing play resulted in a large change in the odds ratio for association between play and milestone attainment, suggesting possible confounding or collinearity between these covariates. Given that play is likely a subset of stimulation time, it is unsurprising that there would be a high degree of collinearity between these variables. For this reason, we discarded stimulation time from the planned multivariate analysis. Based on the analysis described above, we selected play and listening to music for inclusion in the core model, as these represented independent and important factors in milestone attainment within this data set.

Creation of this model was based on the following core assumptions. First, we combined all study groups as it was assumed that the association between each stimulation measure and child development would be stable across arms and time points, and combining study groups allowed us to increase our power through a larger N. Second, the ECCD instrument did not include any questions that permitted differentiation of psychosocial stimulation that was or was not a result of the WV intervention. Hence, there was uncertainty about the place of psychosocial stimulation in the causal pathway between the WV intervention and milestone attainment. For the initial analyses, we elected to assume that most psychosocial stimulation would not lie in the causal pathway.

Phase 6: determining potential confounding covariates

The goal of phase six was to determine which of the covariates describing maternal, household, or child characteristics should be considered for inclusion in the multivariate analysis as potential confounders of associations between the WV intervention and changes in developmental outcomes. This was addressed through a series of modified bivariate analyses (modified in that child age and design variables were included, as previously described) investigating the association between each covariate of interest and milestone attainment in a single model including all four study groups (defined by study arm and time). The *a priori* decision to include all study groups (vs. an analysis stratified by arm and time) was based on the assumption that the association between each covariate and child development would be stable across study arm and time and the desire to preserve the larger sample sizes (as noted above). Only three covariates demonstrated marginally statistically significant associations with milestone attainment (defined as a p value <0.1): dietary diversity, diet containing iron-rich foods, and maternal depression. For a complete list of covariates explored, see Table 3.

Phase 7: construction of the multivariate model

Phase seven was anticipated to be the final stage of analysis, and consisted of the creation of the multivariate model combining the core model developed in phase five, with potentially confounding covariates identified in phase six and an interaction term for time and arm (to permit calculation of difference-of-differences). To determine which combination of covariates represented the optimum set for inclusion in the multivariate

model, we added maternal depression, dietary diversity, and iron-rich foods to the core model (that included self-reported play, listening to music, child age and design variables) in a stepwise fashion, similar to the process undertaken in phase five. As before, a change in an odds ratio of greater than 10% was used to suggest the presence of a complicating factor, such as confounding, collinearity, or presence in the causal pathway.

With the addition of maternal depression to the core model, "play," listening to music and maternal depression all maintained their independent associations with milestone attainment, suggesting that each covariate may have impacted child development through separate pathways. Regarding nutrition measures, there was a high degree of collinearity between dietary diversity and iron-rich foods (unsurprising given the manner in which each variable was constructed, see Table 1), making the inclusion of both nutrition variables in the final multivariate model inappropriate. Given that all covariates maintained independent associations with milestone attainment in a model consisting of the core model, maternal depression and iron-rich foods, we determined that this combination of covariates represented the most appropriate combination for the final multivariate model.

The final step in analysis was to use a series of lincom ("linear combinations of estimators") commands to compare the magnitude of the change in milestone attainment in the intervention community between time points to that occurring in the comparison community between time points. This difference-of-differences analysis allowed us to address the primary study question, regarding whether the WV intervention had resulted in a measurable improvement in child development. It was believed that this model would provide the greatest insight into the role of psychosocial stimulation in child development within this dataset. However, as will be discussed at detail in the following section, unanticipated findings in our preplanned analyses necessitated further exploratory analyses to help us understand the limitations of this final multivariate model.

Results

Descriptive analysis: baseline data

As noted above, descriptive studies comparing the intervention and comparison communities at baseline were largely completed by Mukerjee.[39] In the prior study, there was no significant variability between intervention and comparison communities in regards to maternal, child, or household demographic and risk factor profiles, nor in milestone attainment, as confirmed by χ^2 and T-testing.

Prior work did not include comparisons of psychosocial stimulation time or activities, thus these baseline comparisons were completed as part of this thesis and did detect some baseline differences across study arm. A significantly higher proportion of respondents reported listening to music with their children in the intervention community compared to the comparison community ($p = 0.02$) at baseline. Similarly, more respondents reported singing to their children in the intervention community at a marginally significant level ($p = 0.07$). In contrast, a higher proportion of parents reported speaking to their children in the comparison than intervention community ($p = 0.01$) at baseline. There was no significant variability between comparison and

intervention communities in parental report of play, reading, massage or stimulation time at baseline (see Table 2).

Description of study participants and households at baseline and endline

Data were available for 396 households with children between 6 and 23.99 months. They were unevenly distributed between study arms and time points: 120 in the comparison arm at baseline, 36 in the comparison arm at endline, 109 in the intervention arm at baseline, and 131 in the intervention arm at endline (see Table 2). As mentioned in the methods section, these sample sizes were substantially smaller than the pre-defined target sample sizes. Table 2 describes the developmental milestone attainment, child stimulation measures, and important maternal, child, and household characteristics of participants in comparison and intervention groups, at baseline and endline. In Table 2 the intervention endline group was further divided by exposure status, defined as self-reported participation in one or more sessions of the WV intervention, with 96 unexposed and 27 (22%) exposed households. Eight exposure-unknown households were excluded from Table 2. Measures of milestone attainment and child stimulation were not adjusted for child age in Table 2, and therefore we elected not to conduct statistical significance testing on the crude data.

Inspection of Table 2 suggests the presence of important differences across time and arm, with patterns that are not always consistent with investigators' expectations. Notable and unexpected trends in Table 2 are described below. First, the proportion of children who attained all milestones decreased in the comparison community from baseline to endline, whereas in the intervention community milestone attainment increased between time points. This decrease in milestone attainment in the comparison community was initially surprising, and led to further consideration of the potential effect of differing child age profiles in the two communities (as older age is strongly associated with greater child development). The median age of the index child was approximately equal in comparison and intervention communities at baseline; however age changed in opposite directions in comparison and intervention communities over time, with a decrease of approximately 1.1 months in the comparison community, and an increase of 1.3 (unexposed) or 0.3 (exposed) in the intervention community. In light of this, it seemed less surprising that milestone attainment decreased in the comparison community as it tracked in the expected direction given the changes in child age. Second, regarding psychosocial stimulation activities, the proportion of households that reported participating in each activity increased from baseline to endline in both comparison and intervention communities for all stimulation activities, but the increase was typically greater in the comparison community relative to the intervention community. This pattern was not anticipated for two reasons: a. the WV intervention, if successful, would have increased psychosocial stimulation in the intervention arm alone; and, b. increased psychosocial stimulation was expected to result in better, not worse, developmental outcomes (as noted in introduction). Finally, a substantially higher proportion of respondents reported feeding their children iron-rich foods in the comparison community at endline relative to all other subgroups, a pattern that was not consistent with the decline in median age and milestone attainment in that subgroup. Given the limitations of Table 2 (small n of each subgroup and lack of age adjustment) it is difficult to interpret the meaning of the above trends; however, the existence of unexpected and differing

trends by time and arm drew into question the validity of data and/or comparability of our comparison community, as will be discussed in greater detail in the discussion section.

In other respects, the two arms were similar: play was the most commonly reported form of psychosocial stimulation, and reading the least common, in both communities and at both time points. Regarding stimulation time, comparison and intervention communities saw substantial increases in the proportion of respondents reporting greater than two hours of stimulation time from baseline to endline. However, had the WV intervention been successful, the observed increase would have been expected to be greater in the intervention arm, which was not observed.

In addition to factors noted above, our ability to draw robust conclusions from any of the numerous comparisons possible in Table 2 is also significantly limited by missing values in several domains (especially psychosocial stimulation), and by the difficulty of identifying statistically significant associations in the face of multiple comparisons.

Below, for completeness, we describe the eight children in the endline intervention community with unknown exposure status, and who were excluded from Table 2. Notable characteristics of these eight children are: a. average age of 9.9 months (lower compared to all other study groups); b. higher proportion of children who attained all their milestones relative to all other groups (85.7%); c. lower proportion of parents who reported playing with their children (50%), and higher proportion who reported listening to music (50%); d. lower proportion of female children (12.5%). Because of the small number of children in this category, apparent differences may be due to chance alone.

Bivariate results

Table 3 illustrates the results of modified bivariate analyses (undertaken in phase six) describing the association between covariates of known importance to child development and milestone attainment (in analyses that were adjusted for child age, questionnaire, arm and time, as discussed in methods). In the "All participants" column of Table 3 (columns headed "Stratified Analyses" will be discussed later) children from both study arms and time points were evaluated as a single group, for reasons noted previously. In the pre-planned "lumped" analyses shown in this column, two covariates were found to be positively and significantly associated with milestone attainment: parental report of play, with an odds ratio of 1.84 (p = 0.011, 95% CI 1.15, 2.93), and dietary diversity, with an odds ratio of 1.72 (p = 0.033, 95% CI 1.04, 2.83). One covariate, maternal depression, demonstrated a significant negative association with milestone attainment with an odds ratio of 0.97 (p = 0.005, 95% CI 0.94, 0.99) for each one-point increase on a 60-point depression scale. Two other factors demonstrated marginally significant positive associations with milestone attainment: listening to music (OR 1.73, p = 0.051, 95% CI 0.997, 3.01) and consumption of iron-rich foods (OR 1.48, p = 0.098, 95% CI 0.93, 2.36). As expected, age of child was strongly associated with milestone attainment with an odds ratio of 1.20 for each one-month increase in child age (p= 0.000, 95% CI 1.12, 1.30). No other candidate covariates demonstrated statistically significant associations with milestone attainment in the modified bivariate analysis containing all participants (see Table 3).

Multivariate results

The pre-planned multivariate model would have included a term for household-level exposure to the WV intervention. However, as mentioned in phase three above, this term was not statistically significant either in planned, modified bivariate (see Table 3) or multivariate analyses (data not shown). This finding suggested either that: a. the WV intervention had not been successful; b. it was successful, but success was based on the community-level, rather than household-level, exposure to the intervention; or, c. the n of exposed households was so small that this variable lacked the necessary power to demonstrate underlying relationships.

Therefore, we dropped the household-level exposure term from the pre-planned multivariate model, and proceeded with a version that contained important exposures, confounders and design variables ("play", listening to music, maternal depression and iron-rich foods with terms for age, questionnaire, and time/arm interaction). The results are represented by model 1a in Table 4. The difference-of-differences analysis, comparing the change in milestone attainment from baseline to endline in the intervention vs. comparison community, suggested that the WV intervention was associated with an odds ratio of 6.94 and markedly large confidence interval (p= 0.005, 95% CI 1.79, 26.88) for change in milestone attainment over time, as compared to the change in the comparison community. However, the odds ratio for change in milestone attainment between baseline and endline within the intervention community alone was only 1.83 (95% CI 0.99, 3.35) and was not quite statistically significant (p = 0.052). Within the comparison community, the odds ratio for meeting all milestones at endline vs. baseline was 0.26 (p = 0.033, 95% CI 0.08, 0.90), suggesting that attainment of child development milestones in the comparison community within this model worsened significantly over time. Although suggested on earlier inspection of Table 2, this apparent deterioration in child development status within the comparison community was an unexpected finding, and in light of the lack of significant improvement in child development within the intervention community, it may have been the primary driver for the apparent significance of the results of the pre-planned difference-of-differences analysis. This pattern violated the assumptions of the difference-of-differences model, which assumed that significant changes in difference-of-differences would be caused primarily by the study intervention.

Given the unanticipated changes in developmental outcomes over time within the comparison community, we were not convinced of the reliability of the "final" multivariate model, and so we embarked on a series of additional exploratory analyses seeking to explain these findings. The main observations, concepts and questions driving the exploratory analyses were the following: a. inspection of Table 2 revealed unexpected and internally inconsistent patterns of the evolution of psychosocial stimulation, milestone attainment and nutrition variables across time and arm (as noted earlier); b. psychosocial stimulation and child nutrition may or may not have lain in the causal pathway from WV intervention to developmental milestone attainment; and, c. the sample size of the comparison community at endline was very small.

Unplanned (exploratory) analyses and results

As discussed above, we originally noted the presence of several unexpected trends by time and arm for covariates in the milestone attainment, psychosocial stimulation and

child nutrition domains in Table 2. Inspection of Table 2 had raised doubts about the consistency of the associations of key variables with developmental outcomes; however, given the challenge of interpreting a table unadjusted for age and the small sample size in the comparison arm at endline, we initially hesitated to ascribe too much significance to these trends. In light of the unexpected findings of model 1a, we revisited Table 3 and embarked on series of disaggregated analyses to better understand the unusual patterns observed in several domains.

First, we expanded Table 3 to include stratified bivariate analyses, in order to evaluate our assumption that odds ratios for associations of milestone attainment and specific study covariates would be consistent across arm and time. This analysis revealed that odds ratios for associations with developmental outcomes and covariates in the psychosocial stimulation domain diverged markedly by arm and time. The divergence was most dramatic with the covariate "play," where there was a strong association between "play" and milestone attainment in the comparison community at baseline (OR 3.78, p= 0.03, 95% CI 1.59, 8.95), but not in the intervention community at baseline (OR= 1.52, p= 0.35, 95% CI 0.64, 3.61) or in the intervention community at endline (OR=0.72, p= 0.46, 95% CI 0.30, 1.73). In the comparison community at endline, it was not possible to use the same model to determine the association between milestone attainment and "play" because, within this subpopulation, there were no children without reported exposure to play who passed all milestones, thus precluding our ability to model this relationship (data not shown). For psychosocial stimulation time, the odds ratios for the association between increasing quantities of stimulation time and milestone attainment increased in a stepwise fashion in the intervention community at both baseline and endline, but no clear association existed in the comparison community, (see Table 3, stratified analyses). Notably, with the exception of "play" in the comparison community at baseline, none of the covariates considered demonstrated a significant association with milestone attainment in the stratified bivariate analysis.

This divergent behavior of associations involving "play" by study arm and time in bivariate analyses was unanticipated, and conflicted with our earlier assumption that the intrinsic association between covariates and developmental outcomes would be consistent across arm and time. In light of this, we revisited our initial multivariate analysis (model 1a), this time stratifying it by study arm, to better understand the behavior of "play" in each community, and to shed light on how this divergent behavior may have impacted the model. We created two new models (models 1c and 1i, see Table 5) investigating changes in milestone attainment over time, stratified by study arm. Model 1c, which included play, listening to music, maternal depression and iron-rich foods, showed the results from the comparison community alone, and revealed that within this group the odds ratio for attaining all milestones at endline vs. baseline was 0.19 (p = 0.017, 95% CI 0.49, 0.74), which seemed to confirm the unexpected worsening of development outcomes in the comparison community first noted in Table 2 and model 1a. In contrast, in model 1i, which contained the same covariates but was restricted to the intervention community, the odds ratio for attaining all milestones at endline vs. baseline was 1.91 (p = 0.039, 95% CI 1.03, 3.54). Importantly, while "listening to music", maternal depression, and iron-rich foods maintained similar odds ratios for associations with developmental outcomes across arms, the odds ratios for the association between "play" and changes in milestone attainment over time were dramatically different across study arms. Within the

comparison community, "play" demonstrated a fourfold greater association with milestone attainment over time (OR =4.34, p = 0.001, 95% CI 1.82, 10.34) in contrast to the intervention community where "play" had no observed association with milestone attainment over time (OR = 1.05, p = 0.87, 95% CI 0.57, 1.95). This finding confirmed what was suggested in the stratified bivariate analyses in Table 3: the estimated odds ratios for association of milestone attainment and "play" differed dramatically by study arm, suggesting a significant interaction between "play" and arm and/or time, and therefore any model that included both study arms and "play" and time would be analytically misleading. The construction of a three-way interaction term encompassing arm, time, and "play" might have lent some clarity to the relationships among the three covariates but would have rendered estimation of "differences-of-differences" quite difficult. This revelation led us to regard the findings of model 1a as invalid, however at this stage the reasons for the instability of models incorporating "play" were unknown.

We then readdressed the multivariate analysis in two main ways: 1. we reconsidered the use of other candidate covariates in the psychosocial stimulation domain in an effort to generate a more valid model of the relationship between psychosocial stimulation and child development; and, 2. we repeated analyses with and without different potential confounding covariates, due to uncertainty about the stability of associations with each covariate across arms, and their place in the causal pathway.

The most promising alternative variable in the psychosocial stimulation domain appeared to be the categorical variable for stimulation time. Several factors led to this decision: a. few alternative candidates existed, given that "listening to music" had lost all statistical significance even in model 1a; and, b. within Table 3 stimulation time appeared promising in the intervention community, though it was difficult to analyze in the comparison community given the small sample size. Thus, with reservations, we replaced model 1a with a model that included psychosocial stimulation time (measured as a categorical variable: no stimulation, less than two hours of stimulation, and greater than two hours of stimulation), maternal depression, and iron-rich foods (see model 2a, Table 5), but dropped "play" and "listening to music." No association was noted between "stimulation time" and child development within model 2a, but as a further check we disaggregated by study arm (creating two additional models, model 2c and 2i, see Table 5) to evaluate possible differences across study arm (as was the case with "play"). Interestingly, while no statistically significant association of stimulation time and milestone attainment was noted, the stepwise increase in odds ratios noted with increasing stimulation time in the intervention community (and less clearly in the comparison community, see model 2i) persisted, suggested the possibility that an underlying association existed that was being missed due to low power.

To further evaluate the relationship between stimulation time and milestone attainment, and to determine if low statistical power – particularly in the comparison community at endline – was serving as a barrier, we created an additional model (alternate model 2, see Table 6). This model was based on models 2c and 2i, but the term for time was omitted, thereby allowing us to avoid the small sample size in the comparison community at endline. There was no statistically significant association between stimulation time and child development within the comparison community; however, in the intervention community both intermediate and higher levels of stimulation time were associated with milestone attainment at a marginally significant

level (see Table 6). Additionally, when comparing model 2 with alternate model 2, we appreciated that odds ratios for associations of milestone attainment and "stimulation time" also appeared to demonstrate divergent behavior by time. This was evidenced by the fact that within the comparison community, the relationship between the highest level of "stimulation time" and milestone attainment attenuated when time was added to the multivariate model, whereas in the intervention community the addition of time to the model strengthened the association between more than two hours of "stimulation time" and milestone attainment.

We did not have access to any further information with which we could conduct further exploratory analyses to understand why association of milestone attainment with "play" and "stimulation time" diverged by arm and time, and why the patterns of change for odds ratios for association with such closely related covariates behaved in opposite manners relative to each other. We anticipated that play activities would be an important contributor to a child's total stimulation time, and therefore we expected these "play" and "stimulation time" to track together in each community; however, for reasons we can not determine, this pattern was not observed within this study.

Thus, we concluded that both primary measures of the psychosocial stimulation domain ("play" and "stimulation time") should be removed from all non-stratified analysis due to divergent and inconsistent behavior, and we revisited the primary study question through the creation of two additional exploratory multivariate models (see Table 5) which excluded all covariates in the psychosocial stimulation domain. Model 3 estimated the change in milestone attainment from baseline to endline in intervention vs. comparison communities in a model including maternal depression and iron-rich foods. In this model there was no significant change in milestone attainment from baseline to endline within the comparison community (OR=0.55, p = 0.195, 95% CI 0.22, 1.36); however, in the intervention community milestone attainment significantly improved between time points (OR = 2.03, p = 0.018, 95% CI 1.13, 3.66). The difference-of-differences analysis revealed that the odds ratio for the association of residence in the WV intervention community and attainment of milestones over time was 3.7 (p = 0.017, 95% CI 1.27, 10.78). These trends of no change in the comparison community, and improvement in the intervention community that appeared to be associated with the WV intervention, were consistent with expected study findings.

Given concerns that maternal depression and iron-rich foods may have been present on the causal pathway between the WV intervention and changes in child development, a fourth model (see Table 5) was created that excluded these covariates. As in model 3, model 4 found no significant change in milestone attainment from baseline to endline within the comparison community (OR=0.93, p =0.87, 95% CI 0.41, 2.11), but did find significant improvement in milestone attainment from baseline to endline within the intervention community (OR= 1.92, p =0.023, 95% CI 1.09, 3.36). However, unlike model 3, within model 4 the difference-of-differences analysis found no statistically significant association between milestone attainment and the WV intervention (OR= 2.05, p = 0.153, 95% CI 0.77, 5.48).

Notably, in models 1a, 2a, and 3, associations of milestone attainment with maternal depression and iron-rich foods were consistent across models (see Table 5). In all models the odds ratio for the association between maternal depression and milestone attainment was approximately 0.97 (p <0.01) for each one-point increase on a 60-point

depression scale, demonstrating a strong negative association between greater severity of maternal depression and child development. Regarding the odds ratio for the association between iron-rich foods and milestone attainment, no model demonstrated a statistically significant association, however in all models the directionality of the association was the same and there was significant overlapping of the 95% confidence intervals, suggesting some reliability of this covariate. The stable odds ratios for the maternal depression and nutrition covariates across models, combined with their impact on the difference-of-differences findings in model 3 vs. model 4, suggests that these covariates behave predominantly as confounders within this data set and do not play a significant role in the causal pathway between the WV intervention and developmental outcomes (see discussion).

Discussion

Introduction: the shifting focus of this thesis

At the outset, this study focused on inferential questions, with twin goals of answering the primary study question ("Was the WV intervention associated with improved developmental outcomes?") and shedding light on the relationship between psychosocial stimulation and child development within this data set. However, as will be discussed at length in the following sections, numerous methodological constraints, related to study design, recruitment, and the ECCD questionnaire, severely limited our ability to make inferences related to pre-specified questions posed for this exercise.

Importantly, as will be illustrated below, the focus of this thesis evolved into one that predominantly highlighted methodological questions in an area of research where several measurement issues are still unresolved. Consequently, much of this discussion will emphasize the methodological constraints with which we struggled during these analyses, as well as recommendations for avoiding similar challenges in future work.

Interpretation: planned and exploratory analyses undertaken to address the primary study question

With regard to the pre-specified primary outcome, based on this data set and these analyses, the answer to the primary study question, "was the WV intervention associated with improvements in child development?" remains elusive. This results from numerous factors, all of which will be discussed in greater detail, including: a. no significant association observed between household-level exposure to the WV intervention and improved developmental outcomes; b. the unexpectedly divergent behavior by arm and time of odds ratios for associations of milestone attainment with covariates in the psychosocial stimulation domain; c. the lack of clarity regarding what factors exist on the causal pathway between the WV intervention and changes in developmental outcomes; and d. numerous methodological constraints including small sample size, questionable comparability of selected intervention and comparison sites, and the inclusion of only two clusters (one cluster per arm).

The initial intent of this study was to include a term for household-level exposure to the WV intervention, as this would have allowed us to measure direct vs. indirect effects of the WV intervention and reduce confusion regarding whether covariates did or did not exist on the causal pathway between the WV intervention and changes in

developmental outcomes. However, very early in these analyses, household-level exposure to the intervention was dropped from all models because there was no significant association with milestone attainment in any modified bivariate or multivariate model. As discussed previously, the reason for this is unclear. It may suggest that the WV intervention was unsuccessful, or it may simply be related to the manner in which exposure was measured in this study, and therefore does not reflect the true impact of the WV program. Regardless of the cause, the fact that household-level exposure to the intervention was dropped from all models meant that the study models included in this paper can only address community-level (not individual-level) effects of the WV intervention, which in turn lessens our ability to suggest a direct causal relationship between intervention and outcome.

As discussed in the results section, creating a multivariate model that accurately reflected the underlying relationships among covariates proved challenging (see Table 5). This was particularly true for covariates in the psychosocial stimulation domain, particularly "play" and "stimulation time." The reason associations between the primary outcome and these two covariates diverged by arm, and furthermore evolved in an opposite manner relative to each other over time, is unclear and is discussed further in a subsequent section. Possible explanations include: a. the nature of psychosocial stimulation was fundamentally different in the two communities; b. parental self-report of stimulation differed by arm and/or time; c. the methods of interviewing regarding stimulation differed between communities; or d. fundamental differences between the communities (unrelated to psychosocial stimulation) resulted in different associations between exposure to psychosocial stimulation and outcome. Regardless of the underlying cause, the unreliability and internal inconsistency of odds ratios for associations with covariates in the psychosocial stimulation domain, combined with a question of whether or not these covariates existed in the causal pathway between the WV intervention and changes in developmental outcomes, necessitated their removal from all multivariate models, as inclusion of these covariates clouded our ability to answer the primary study question.

Both exploratory model 3 and model 4 excluded all psychosocial stimulation covariates for reasons given above. In order to determine which of these two models more accurately illustrated the impact of the WV intervention, it was necessary to determine which model was a more appropriate fit for this data set. Model 3 (which controlled for maternal depression and iron-rich foods) proved a better model for the following reasons: a. examination of changes in odds ratios for the difference-of-differences findings across the two models suggested that maternal depression and iron-rich foods acted predominantly as confounders within this data set, and did not exist on the causal pathway between the WV intervention and developmental outcomes (discussed in more detail below); and, b. as discussed in the background section, ample evidence exists in the literature indicating that maternal depression and child nutritional status can confound the relationship between various exposures and child development.

We therefore elected to endorse model 3 as the most appropriate model to answer our primary study question, despite the fact that it was a *post hoc* analysis. The difference-of-differences analysis in this model indicated that residence in the intervention community was associated with a statistically significant improvement in milestone attainment, suggesting that the WV intervention resulted in improved

developmental outcomes through a community level mechanism. That said, the parent study suffered from numerous methodological constraints (discussed below), which hamper our ability to infer that improvements in developmental outcome resulted from the WV intervention, as opposed to reflecting underlying and unmeasured confounding factors. Therefore while we believe our analyses suggest that the WV intervention likely resulted in improved developmental outcomes, we are unable to definitively prove that this is the case, nor are we able to demonstrate the causal pathway through which these likely improvements occurred.

Methodological constraints: selection of study communities and households

There are several important methodological constraints associated with the design and implementation of the parent study that impact our interpretation of study findings. Constraints of particular concern given their threat to drawing inferential conclusions include: a. innate imbalances between the comparison and intervention communities that result from poor site matching; b. the potentially modest "dose" associated with the WV intervention; c. apparent differences in characters of the comparison and intervention communities at both baseline and (to an even greater extent) endline; d. the small sample size and different recruitment practices undertaken in the comparison community at endline; and, e. the inclusion of only one cluster (community) per arm. These constraints may have lead to a fundamental non-comparability of study communities, and will be described in more detail below.

Baseline imbalance between the comparison and intervention communities that result from the lack of strategic matching of study sites is a significant constraint to analysis, as it has the potential to introduce numerous confounding elements. The WV study protocol, written before the study commenced, discussed the importance of intentionally selecting comparison and intervention communities that were as similar as possible in order to minimize potential confounding. However, given financial and logistic limitations, WV had little flexibility in the choice of study locations. Ultimately, the study included the only two locations in Tijuana where WV had an established presence, and the comparison and intervention communities differed even at baseline (endline differences will be discussed below). Significant baseline differences included psychosocial stimulation practices, where it was noted that "play" was strongly associated with improved developmental outcomes in the comparison community only, and "stimulation time" was marginally associated with improved outcomes in the intervention community only. These were key covariates in our analyses, and because this study only included one community/cluster per arm, we were unable to account for these baseline differences through the mechanisms typically employed in cluster studies that include numerous communities/clusters per arm. Therefore, the inherent differences across arms at baseline made it difficult to determine if changes over time were the result of the WV intervention, or instead represented unmeasured differences between the communities.[49,50] Given this concern, study findings should be interpreted with caution.

Another important limitations relates to the potentially modest "dose" delivered by the WV intervention. This "dose" was not well documented, as it was based on unsystematic delivery of the curriculum as designed, and the relatively brief period between measurement points. Because of these factors, the "dose," or potentially beneficial effect of the WV intervention on a population level, could be small.

Additional methodological constraints related to peculiarities of the comparison community include the following: a. at endline the comparison community seems to be fundamentally different from all other subgroups, including the comparison community at baseline; and, b. the evolution of baseline and endline communities over time is unexpectedly different. This can be appreciated in Table 2 in the following ways: a. the proportion of children who attain all their milestones unexpectedly decreases to an extent that is not explained by small differences in median age alone; b. the proportion of parents who report performing each of the psychosocial stimulation activities increases more over time in the comparison community relative to the intervention community (contrary to what would be expected had the WV intervention been successful); and, c. nutrition measures also change more in comparison than intervention communities and do not track in parallel with the expected age effect (the proportion of children receiving and iron-rich diets increases dramatically in the comparison community at endline despite the decrease in median age).

We considered several possible explanations for the observed differences across study subgroups. First, the n of the comparison community at endline is approximately one third the size of all other sub groups (n=36), and barely 10% of the intended sample size. If this very small sample were representative of the characteristics of the comparison community at endline, sample size alone would reduce statistical power but should not result in the marked changes in odds ratios for associations of key covariates with developmental outcomes that we observed. That said, the small (and occasionally zero) cell sizes that existed in several of the subcategories of the stratified bivariate analyses did result in undetermined odds ratios, complicating interpretation of these analyses. Therefore while small sample size alone should not introduce bias into our study sample, it may have contributed to the unexpected difficulties in interpretation of data.

A second arguably more concerning explanation for the fundamental differences of the comparison community at endline is the unplanned alteration of recruitment practices. Unlike all other groups, modest financial incentives were provided for study participation in the comparison community at endline as a means to quickly address resistance to participation. This practice likely introduced bias through several mechanisms. Participants in the comparison community at endline were motivated to enroll for fundamentally different reasons from those in all other study groups, and therefore participants in this group likely differed in important ways. Additionally, paying participants to answer survey questions changes the relationship between the interviewer and the participant. For example, in interviews where individuals were paid for their participation, participants may have felt increased social or financial pressure to provide what they felt were the preferred answers (social desirability bias). If this were to have been the case, it might explain the greater increase in the reported levels of psychosocial stimulation in the comparison community, relative to the intervention community, over time.

A final possibility for the unexpected trends in the comparison community by arm and time, is that the composition of this community changed over time in ways that could not be detected with the survey instrument. Given the limitations of study design we are unable to evaluate this possibility at this time.

Regardless of the cause of the differences appreciated in the comparison community at both time points and over time, these differences were a significant

constraint to analysis as they meant that the comparison community at endline likely did not provide an ideal control group for either the comparison community at baseline or the intervention community. As a result, the unanticipated changes in odds ratios for key associations when analyses were disaggregated by arm and/or time, as well as the discordant difference-of-differences conclusions we reach through comparing different models, may reflect these fundamental differences in study groups by arm and time and not the true associations we were seeking to measure. Given the limitations of the ECCD questionnaire, and the fact that this study only included one cluster per arm, we are unable fully assess the degree to which non-comparability contributes to study findings, and therefore this factor threatens the validity of study conclusions.

Exploratory analyses: understanding relationships in the psychosocial stimulation domain and implications for future studies

Fundamental differences across study arms and time points may be the primary reason for the unexpected and inconsistent behavior of key covariates in the domains of psychosocial stimulation and child development, as noted above. But it is important to consider other possible reasons why odds ratios for associations of the primary outcome with the covariates "play" and "stimulation time" diverged, and with opposite directionality relative to each other. We were unable to identify any logical pattern that would be expected of variables that truly represent underlying associations, and therefore we hesitate to draw any conclusions about the overall relationship between psychosocial stimulation and child development within this data set. What follows is a discussion of several potential explanations for the divergence (other than non-comparability of study subgroups by arm and time), and our recommendations section discusses different strategies that could be employed in future studies to avoid these concerns. This discussion will focus on the covariate "play," as the odds ratio for its association with milestone attainment differed more dramatically between groups (defined by arm and time). Similar factors likely affected observed associations with "stimulation time," but here we will focus on "play" alone.

It is unclear if the nature of the play itself, parental self-reporting of play, and/or methods of interviewing regarding play led to the observed inconsistencies. Evaluating the nature of "play" in each community is impossible given the constraints of the questionnaire. As a dichotomous variable, "play" simply indicates any amount of play vs. none over an unspecified time period, and does not provide any information regarding the quality or quantity of play. Therefore, within this data set, we have no understanding of the character of the playful interactions that took place, and can't shed light on the question of whether or not the nature of "play" was fundamentally different between the two study groups or over time. The unexpected instability of odds ratios for associations of milestone attainment with "play" may also have resulted from different patterns of parental reporting of "play" by study arm and over time. In studies such as this, the most concerning possibility would be differential misclassification as a result of biased reporting within the exposed group. However, this particular form of differential misclassification is unlikely in this study for the following reasons: a. there was no association between "play" and milestone attainment in the intervention community at baseline or endline; and, b. the overall increase in parental report of play was greater in the comparison community (the increase would be expected to be greater in the exposed

intervention community). Therefore, while we can't rule out differential parental reporting, it is unlikely to have resulted from exposure to the intervention. It is possible that there were unintended and unrecorded differences in the interviewing methods employed in each study arm. It is impossible to retroactively determine if differences in interviewing techniques took place, but it is notable that the same team of interviewers and field supervisor was deployed in both communities, but differed between the two time points. Finally, it is also important to note that different secular trends (such as changes in policies, healthcare availability, or population composition) between comparison and intervention communities may also have affected psychosocial stimulation practices or reporting, and that this possibility can't be controlled for within the constraints of this study given that it included only one cluster per arm.

During the exploratory analysis phase of this thesis work, we have come to the realization that some of our inability to evaluate and understand the true associations of "play" with developmental milestone attainment is rooted in the limitations of the ECCD questionnaire, including lack of: a. clear definitions of key measurement constructs and the questions used to measure them; b. questions that would permit checking the consistency of information and thus the internal validity of the questionnaire; and, c. consideration of the likely causal pathway between the WV intervention, changes in play behavior, and changes in milestone attainment during the design of the questionnaire. These are described in greater detail below, and recommendations for future study design is included in a later section.

The ECCD questionnaire did not clearly define what was meant by the term "play." As different people and cultures likely define play in different ways, there is a high degree of uncertainty regarding what the covariate "play" truly represents in this study. This lack of clarity could have resulted in the nature of play, or reporting of play, differing between study arms and over time, thereby contributing to the unexpected findings outlined above.

While the ECCD questionnaire included multiple questions regarding psychosocial stimulation, the instrument provided limited opportunities for us to verify consistency or internal validity of covariates. For example, though the questionnaire included questions about stimulation activities and total amount of stimulation time, due to limitations in design, we were not able to deduce how much stimulation time a child received in each of the separate activities, or which activities were undertaken during the specified period of time. Thus the role of "play" as a contributor to "stimulation time" is unknown.

Finally, to our knowledge, the causal pathway through which the WV intervention was expected to generate change in play behaviors was not fully articulated by the primary investigators designing the parent study. However, the importance of anticipating the causal pathway can't be underestimated, and in this study would have been an important consideration during both questionnaire development and design of the analytic plan (discussed below). Because the anticipated causal pathway was not pre-specified, the ECCD questionnaire did not necessarily include questions that would get at the elements of "play" that were most likely to impact developmental outcomes, or most likely to be influenced by the WV intervention. For example, it is conceivable that the WV intervention was more likely to help parents better understand what types of playful interactions most benefit child development, than it was to motivate parents who

typically do not play with their children to begin playing. If this were to be the case, then the dichotomous covariate "play" would not capture the meaningful changes generated by the WV intervention, and would therefore fail to demonstrate underlying associations that truly existed. While this is simply a hypothetical example, it illustrates the potential implications of not considering the likely causal pathway from the outset.

Covariate roles: presence in the causal pathway vs. confounding

As mentioned above, consideration of the anticipated causal pathway between intervention and outcome is essential during the design of the analytic plan. The significance of this step is highlighted by the challenge we faced in our exploratory analyses that sought to determine which multivariate model best fit our data set. Through this process we recognized that it was unclear whether several of our covariates were confounding the relationship between the WV intervention and improved developmental outcomes, existed on the causal pathway between the WV intervention and improved developmental outcomes, or were a hybrid of these two possibilities. This consideration affected maternal depression and iron-rich foods, as well as covariates in the psychosocial stimulation domain.

As mentioned previously, we ultimately concluded that maternal depression and iron rich foods acted predominantly as confounders and did not exist on the causal pathway, making model 3 (see Table 5) the best fit for this data set. As discussed in the background section, the strength of the literature highlighting the impact of maternal depression and iron-rich foods on child development is impressive, and this alone argues in favor of including these two covariates as potential confounders. This view is reinforced by the stability of odds ratios for associations of the primary outcome with these covariates in this data set, where we see that across both arm and time, greater severity of maternal depression is associated with worse milestone attainment at a highly significant level, and while not significant, iron-rich foods maintains a consistent positive association with improved outcomes. Importantly, comparison of models 3 (with maternal depression and iron-rich foods) and model 4 (without these covariates) further demonstrates that these two covariates exist predominantly as confounding factors, as supported by the decrease in odds ratio for the difference-of-differences analysis by over 10% when maternal depression and iron-rich foods are excluded from the model.

As discussed in the results section, during the exploratory analyses, it became clear that covariates in the psychosocial stimulation domain had to be removed from all multivariate models. There were two key reasons for this, including the puzzling changes in odds ratios for association of outcomes with these covariates by arm and time (discussed previously), and ambiguity regarding whether or not these covariates existed on the causal pathway between the WV intervention and changes in developmental outcomes. Theoretically, since the WV intervention discussed the importance of psychosocial stimulation to child development, the covariates "play" and "stimulation time" could have existed on the causal pathway between the WV intervention and milestone attainment. Due to the challenge of interpreting instability in odds ratios for association of milestone attainment with these two covariates, we were unable to analytically evaluate the degree to which these covariates acted as confounders vs. resided in the causal pathway, as we did with maternal depression and iron-rich foods. Therefore, given the unresolvable ambiguity, it was determined that models excluding

psychosocial stimulation covariates provided more appropriate representations of study relationships.

It is also worth noting that these problems were exacerbated both by our inability to control for exposure to the intervention, as well as the fact that it is unknown at this time what educational material was ultimately presented during the WV program. As discussed previously, we were unable to generate a reliable variable denoting exposure to the WV intervention, due to the large amount of missing data and small number of participants reporting exposure to the WV intervention. Without an exposure variable, we were unable to control for effects of direct participation in the WV program in our multivariate models, and therefore could not use our models to help determine what did and did not exist on the causal pathway.

Furthermore, we were unable to hypothesize about what covariates were more or less likely to be impacted by the WV program, because the true content of the WV program as actually delivered is unknown. Instead of following the planned curriculum, WV allowed session topics to be dictated by group demand and the final agenda went unrecorded. This ambiguity further decreases our ability to evaluate whether a covariate was likely to exist on the causal pathway or act as a confounder because we have no way of knowing which factors were more or less emphasized during the program. However, we are aware that maternal depression was not an explicit component of the planned curriculum.

Challenges inherent in measuring child development

The majority of this discussion has focused on methodological constraints, including flaws in the WV parent study design, limitations of the ECCD measurement tool, and small sample size, as these factors have hampered our ability to answer our primary inferential questions. An additional methodological issue is the interpretation of the primary outcome measure, milestone attainment. Below we include a brief discussion of the constraints associated with milestone attainment as an outcome measure in this study.

Milestone attainment, as measured in this study, may not be an accurate reflection of a child's developmental status for three key reasons: a. bias towards ascertainment of the development of motor skills; b. imbalances in milestones across age categories; and, c. paucity of milestones assessed. The ECCD questionnaire contained unequal numbers of questions assessing each developmental domain: six gross motor milestones, three language milestones, one social/emotional milestone and two additional milestones that did not reflect a specific domain. No milestones assessed in the ECCD questionnaire fell into the domain of fine motor development. As a result the variable passed-all-milestones is biased towards the development of gross motor skills, and may not accurately reflect a child's cognitive or social/emotional well-being. Additionally, though each ECCD questionnaire was administered to children across a range of ages (either 6-11.99 months or 12-23.99 months), questionnaires did not include adequate numbers of milestones assessing developmental status in each age bracket subgroup of interest. For example, the questionnaire administered to children of 12-23.99 months did not include any milestones that would typically be achieved after about 15 months of age. The impact that the poor distribution of questions across the age range will have on a child's likelihood of passing all milestones depends on the age of the child, making it an unequal measure across age

brackets. While the variable "milestone attainment" as measured in this study provides a general sense of a child developmental status, it is clear that a more accurate representation would be obtained from a measurement tool that included a set of questions that were balanced across developmental domains and age ranges. Additionally, consideration of whether the assessment tool is applicable across different cultural settings is essential, as will be discussed in greater detail below.

Study conclusions and recommendations

As discussed in the background section, the impact that early child development programs can have on a child's long term health outcomes is now widely recognized, and as a result the health community has seen a dramatic increase in interest in investing in early childhood interventions worldwide.[1,2,51] In light of this, we will conclude by highlighting some important lessons learned from this analysis. We will focus on four key conclusions from this study, specifically: a. causal pathways between exposures and outcomes must be considered during study design; b. exposure and outcome covariates need to accurately and reliably reflect important elements of the concepts of interest; c. robust study design is essential to enabling a meaningful analysis; and, d. maternal depression is strongly associated with worse developmental outcomes. Each of these concepts, and related recommendations, will be discussed in greater below.

1. Importance of considering the likely causal pathway during study design

As discussed above, it is difficult to overstate the importance of considering the likely causal pathway between an intervention and anticipated outcomes during the design of the study protocol and analytic plan. While this philosophy applies to all scientific research, it is particularly important in the arena of child development, where interventions, such as the one undertaken by WV, are often broad in nature and may cover everything from discipline practices, to hygiene, to child nutrition. In this setting, the causal pathways between each intervention and the expected changes in parenting behavior or child development outcomes can be numerous and complex, often involving multiple interconnected mechanisms. Articulating the expected causal pathways *a priori* enables investigators to insure that the measurement tools used to evaluate study outcomes are consciously constructed in order to include key concepts that are relevant to the anticipated causal pathways. This increases the overall likelihood that a given measurement tool will be able to capture anticipated changes or underlying associations.

Additionally, articulating complex and interconnected causal pathways *a priori* allows investigators to anticipate potential challenges that may arise during the analytic phase and take any steps necessary to mitigate their effects. The pitfall of not considering these complexities is highlighted by the challenge we faced in determining the causal pathways connecting psychosocial stimulation, the WV intervention, and developmental outcomes (as described in the discussion section). Even in the best-designed study, there will always be a degree of difficulty in the evaluation of early childhood development programs. This is because many of the factors these programs seek to improve, and that therefore by definition exist on the causal pathway between intervention and outcome (such as home environment, child nutritional status, psychosocial stimulation, and maternal mental health), are, simultaneously, powerful confounders of the associations in

question. This inherently complicated dynamic highlights the importance of being able to control for direct exposure to the intervention. Had the WV study been better able to construct a valid exposure variable (or set of exposure variables), it would have been more able to separate and to measure both direct and indirect effects of the intervention, thereby shedding light on the degree to which various covariates acted as confounders vs. existed on the causal pathway. Alternatively, had the WV study been constructed to allow index children to be followed longitudinally (in contrast to a repeat cross-sectional design) parents of index children could conceivably have been linked by name to the attendance lists for various sessions, thereby shedding light on the impact of exposure to specific topics. In studies with this degree of complexity in the causal pathways in question, it is essential to both articulate the likely causal pathways *a priori*, and also construct the study design and measurement tools such that exposure to the intervention can be accounted for in the analytic model.

In addition to anticipating the likely causal pathways resulting from a pre-specified intervention, it is essential to clearly record adherence to, or deviations from, the intended program or the intervention. This allows those evaluating the intervention to better assess likely causal pathways after the fact, and also enables investigators to reconsider potential outcomes if an intervention deviates dramatically from its intended program. As mentioned above, in an ideal setting investigators would be able to reduce confusion regarding the causal pathways of an intervention through the inclusion of an exposure variable in the model. However, as was the case in this analysis, research is imperfect, and in situations where exposure can't be fairly accounted for, knowledge of what topics were emphasized, de-emphasized or omitted can be immensely helpful. For example, it can be inferred that topics strongly emphasized in a program's curriculum should be more likely to exist on the causal pathways between intervention and outcome, and topics that are not covered, by definition, could not exist on the causal pathway between intervention and outcome. Additionally, if the true content of a program differs from the planned intervention to a large degree, it is possible the study outcomes, and causal pathways leading to them, could be completely different from what was anticipated *a priori* (and as a result, may not be measurable by original study instruments). In this setting, clear records regarding what topics were ultimately covered are essential to accurate interpretation of study findings. While greater clarity regarding program content would have been particularly helpful in this study, where causal pathways between the intervention and outcomes were complex and ambiguous, all studies, no matter how well constructed, benefit from diligent record keeping regarding adherence to planned protocols.

2. Measuring exposure and outcome

The second key conclusion regards the importance of clearly defining key exposure and outcome measures. In the discussion section we addressed the challenges we faced as a result of the ambiguous definitions of covariates in the psychosocial stimulation domain (a key exposure measure), as well as limitations that grew out of the imperfect covariate denoting milestone attainment (a key outcome measure). Below we will provide several recommendations for how to avoid similar difficulties in future studies.

As highlighted in the discussion section, many of the limitations associated with covariates in the psychosocial stimulation domain grew out of their ambiguous definitions. We propose three recommendations based on the challenges we had with the covariate "play," including: a. design measurement tools to focus on the aspects of "parental play with children" that are anticipated to be driving any underlying associations with milestone attainment; b. clearly define exposures of interest; and, c. explore exposure covariates through parallel and complimentary questions that would permit internal verification. Insuring that measurement tools focus on the elements of an exposure most likely to drive the desired change will maximize a study's likelihood of demonstrating underlying associations (if they exist) and also increase confidence in negative findings. For example, if a covariate does not capture the element of psychosocial stimulation that drives changes in child development, then negative findings may not indicate that there truly is no underlying association, but rather simply reflect the overall weakness of the covariate in question. Greater clarity in exposure definitions will have several important effects, both within a given study and in future investigations. With clear exposure definitions, a study's internal consistency will improve, especially in a setting where multiple individuals are conducting interviews. Importantly, clearer exposure definitions will also increase the reproducibility of study findings in future investigations and facilitate easier implementation of key conclusions. For example without knowing what was meant by "play," we would have been unable to provide substantive recommendations related to this construct, had this study shown significant associations between "play" and child development. Finally, exploring key exposures through parallel and complementary questions will both allow for internal verification of study findings, and will be helpful in settings where covariates behave in unexpected ways, as was the case with "play" in this study. Had we been able to compare different measures of psychosocial stimulation in the ECCD questionnaire as an internal validity check, we might have been better able to evaluate and interpret the divergent behavior of "play." These three recommendations are by no means an exhaustive list of considerations that should be met to develop strong exposure measures, however they would avoid what we believe to be three key limitations in this study.

As mentioned previously, this discussion has not emphasized the strengths and limitations of the outcome measure "milestone attainment" as we have moved away from a focus on inferential questions. However, there are several key lessons learned in this study regarding both the limitations of our specific outcome measure, as well as the challenges inherent in measuring child development regardless of setting. As mentioned in the discussion section, the covariate for milestone attainment provided a skewed overall view of a child's developmental status, as the milestones assessed in the ECCD questionnaire overemphasized gross motor development and failed to include milestones that would have captured short-term developmental changes or changes in older children. An optimal child development assessment tool would include an equal number of questions evaluating development in each of the four developmental domains: gross motor, fine motor, language, and social-emotional, each of which would need to be balanced across age categories. However, it is notable that creating a measurement tool that provides a robust measure of development for all ages, and can be administered in brief interviews with caregivers, is challenging. In this setting, all data must be gathered from caregiver self-report (thus milestones that can only be appreciated by a trained

observer must be omitted), the overall number of milestones evaluated is limited by time, biometric data (such as blood samples) can't always be obtained, and certain developmental domains tend to be easier to measure and thus are overemphasized. Developing a balanced and robust brief questionnaire assessing child development that can be applied across cultures is currently an active area of research,[31–33] and this study highlights both the importance of, and challenge associated with, developing such a measurement tool.

In light of this, we want to comment on two additional factors innate to child development that pose challenges to the development of a brief assessment tool based on milestone attainment: a. long normal developmental windows for attaining various milestones; and b. difficulties measuring developmental changes over short time periods and across groups. Children are expected to meet certain developmental milestones within a developmental window, with normal development occurring at any point during that time. Normal developmental windows are often long, resulting in broad confidence intervals around the expected child age for attainment of each milestone. For example, according to the WHO child development standards, a child is expected to "walk independently" between 8.2 and 17.6 months, with normal development defined as attaining this milestone at any point during this nine-month period.[45] In this setting, obtaining enough children who fall outside the normal development window can be difficult, and therefore milestones with broad confidence intervals around the range of age of expected attainment may require a large sample size, especially if the outcome measure is defined as a dichotomous "pass" vs. "fail" as opposed to the continuous variable "age at milestone attainment." Another challenge faced by child development investigators relates to the fact that child development studies often take place over short periods of time due to financial and logistic constraints, however many developmental changes only become apparent after longer time periods have elapsed. For example, psychosocial stimulation might be expected to impact emotional and cognitive developmental; however, the impact of these changes may not be apparent until adulthood when behavioral and academic measures can more easily be attained. This trend was recently seen in a 20-year longitudinal study in Jamaica, where behavioral improvements were seen in children included in the psychosocial stimulation arm of a study in long-term but not short-term analyses.[26,27,52] These two time-related challenges are innate to child development and therefore factor into all child development studies. We draw attention to them here to highlight the importance of adequate sample sizes in studies assessing child development via milestone markers, and also the significant benefits that can be obtained from following children longitudinally through time to assess long-term benefits.

An additional challenge to child development research, particularly in resource-limited settings where the burden is highest in regards to children not attaining their developmental potential,[1] is the dearth of child development assessment tools that have been validated across cultures.[38] The majority of child development scales have been developed and validated in western or high resource communities, and as a result may not accurately assess developmental outcomes in resource-limited communities where children may have less access to certain items, or less exposure to specific concepts, built into the framework of the assessment tools.[23] Numerous studies have been undertaken to either validate existing tools, adapt existing tools to new contexts,[31] or develop new child

development measurement tools,[32,33] in low-income communities around the globe. However, even when a tool is developed specifically for use in a resource-limited setting, it is often challenging to know whether or not it will apply across different cultural contexts. This is based on the simple fact that the experience of a child raised in Africa likely differs in fundamental and important ways from that of a child raised in South America or SE Asia. For this reason, a proposed assessment tool would ideally be tested in a small pilot program within the population of interest prior to initiation of the study, however the added time and cost burden associated with this step can serve as a barrier. Therefore, given the reality of research in resource-limited settings, it is essential to be cognizant of the limitations associated with interpreting child development assessment tools outside of their original setting, and to select a tool validated within a similar cultural context.

3. Nuts and bolts of study design

The third important lesson illustrated by this analysis regards the importance of insuring robust nuts and bolts of study design. As discussed previously, the strength of this analysis, and our ability to draw inferential conclusions, was severely limited by problems arising at the level of study design and implementation, namely small sample sizes, the inclusion of only one cluster per arm, and the likely incomparability of the comparison community by arm and time. Recommendations for future studies are straight forward: a. estimate and recruit adequate sample sizes, as this both assures statistical power and is particularly important in demonstrating changes in developmental outcomes (as discussed above); and, b. include numerous clusters per arm, as this allows for more robust analysis, removes the question of whether impacts are simply the result of differences between clusters, and decreases the likelihood of fundamental non-comparability between communities.[43] While these recommendations seem simple, financial or logistic limitations often interfere with the implementation of ideal study designs,[41] as was the case with the WV intervention where actual sample sizes were dramatically less than those obtained in sample size calculations. When investigators are unable to achieve target sample sizes, or reconsider desired study designs, the strength of the study and its capacity to address the primary study question can be significantly decreased. This study provides an excellent example of how severely impaired a study can become when the design is compromised in key arenas.

4. Inferential conclusions: importance of maternal depression to child development

The fourth significant conclusion from this study regards the strong negative relationship between increasing maternal depression scores and developmental outcomes, and is the inferential conclusion from this work in which we have the most confidence. Maternal depression demonstrated a strong negative association with milestone attainment in all study models, a finding that is consistent with trends shown in the published literature.[53–55] It is notable that in this analysis, the variable for maternal depression represented a linearly measured numeric score on a 60-point depression scale. The significant association between maternal depression and decreased milestone attainment reflected the impact of a single point increase on that developmental scale. This suggests that more profound maternal depression may have an even greater negative impact on child development than mild depression, and that even small incremental

decreases in maternal mental health status may lead to worse development outcomes. This indicates that in regard to child development maternal depression should not be viewed in a binary manner, a finding that potentially has significant implications clinically. It suggests that even without relieving maternal depression, small improvements in maternal mental health could represent important clinical targets, as they could result in significant improvements in development outcomes for children. Further investigation is warranted to validate this finding and determine its clinical significance.

It is also worth noting that the maternal depression scale used in this study was the only aspect of the questionnaire that was previously validated. All other portions of the ECCD questionnaire were developed by WV for this study, and were not validated in other settings. The fact that the previously validated portion of the questionnaire was the only element that enabled inferential conclusions to be drawn further highlights the importance of robust measurement tools.

Conclusion

Enthusiasm and funding for programs seeking to improve early childhood development are increasing worldwide, as it is now recognized that interventions at a young age can result in improvements in long-term health outcomes. [1,2,51] Despite this, lack of consensus regarding how best to assess a child's developmental status across cultures is a significant barrier to accurate program evaluation and cross study comparisons.[23,38] Given this dynamic, it is becoming increasingly important to develop robust tools to enable the public health community to assess the efficacy of the child development interventions.

Despite, or arguably because of, the limitations of this analysis, this study provides important insights into the challenges faced by child development researchers. The limitations of the ECCD questionnaire, weak design of the parent study, and complexities associated with assessing child development, highlight important challenges present in many studies evaluating early childhood development programs. All of these challenges are common and may be difficult to avoid in the field experiment, but our hope is that this work will enable future investigations to be better constructed so as to avoid common pitfalls, enabling more robust and impactful studies.

Acknowledgements

Annette Ghee proposed the community-based survey component of the original WV evaluation. The Tijuana-based team responsible for implementing the intervention and the accompanying evaluation was led by Dr. Osvaldo Benitez, WV Latin America and Caribbean Regional Advisor for Research and M&E, and Lic. Noé Martínez, Regional Director for WV México's northern border region. Lic. Patricia Hartasanchez, serving as WV Latin America and Caribbean Regional Advisor for Life Skills and ECD, also took a more active role to support the completion of the project. Dr. Rufino Mechaca, MD, MPH, and Professor of Epidemiology at the Graduate School of Medicine and Psychology at the Universidad Autónoma de Baja California was contracted to lead the data collection and conduct primary analyses. Dr. Osvaldo Benitez and Lic. Beatriz

Alfaro Trujillo, MPH, PhD(c) served as field coordinators for baseline and endline data collection respectively. Lic. Alfaro also collaborated with Dr. Menchaca to prepare the WV final evaluation report using this data set.

Literature Cited

1. Grantham-McGregor SM, Cheung YB, Cueto S, et al. Developmental potential in the first 5 years for children in developing countries. *Lancet*. 2007; 369: 60–70.

2. Walker SP, Wachs TD, Grantham-McGregor SM, et al. Inequality in early childhood: risk and protective factors for early child development. *Lancet*. 2011; 378: 1325–1338.

3. Gertler P, Heckman J, Pinto R, et al. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*. 2014; 344: 998–1001.

4. Walker SP, Chang SM, Powell CA, Baker-Henningham H. Building Human Capacity through Early Childhood Intervention: The Child Development Research Programme at the Tropical Medicine Research Institute, The University of the West Indies, Kingston, Jamaica. *West Indian Med J*. 2012; 61: 316–322.

5. Engle PL, Fernald LCH, Alderman H, et al. Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries. *Lancet*. 2011; 378: 1339–1353.

6. CSDH (2008). *Closing the gap in a generation:health equity through action on the social determinants of health. Final Report of the Commission on Social Determinants of Health.* Geneva, World Health Organization.

7. Chan M. Linking child survival and child development for health, equity, and sustainable development. *Lancet*. 2013; 381: 1514–1515.

8. WHO Multicentre Growth Reference Study Group. Assessment of sex differences and heterogeneity in motor milestone attainment among populations in the WHO Multicentre Growth Reference Study. *Acta Paediatr*. 2006: 66–75.

9. Fox SE, Levitt P, Nelson CA. How the timing and quality of early experiences influence the development of brain architecture. *Child Development*. 2010; 81: 28–40.

10. Britto PR, Pérez-Escamilla R. No second chances? Early critical periods in human development. *Social Science & Medicine*. 2013; 97: 238–40.

11. Heckman J. The economics, technology, and neuroscience of human capability formation. *PNAS*. 2007; 104: 13250–13255.

12. Maggi S, Irwin LJ, Siddiqi A, Hertzman C. The social determinants of early child development: an overview. *Journal Paediatrics Child Health*. 2010; 46: 627–635.

13. Maulik PK, Darmstadt GL. Community-based interventions to optimize early childhood development in low resource settings. *Journal of Perinatology*. 2009; 29: 531–542.

14. Hamadani JD, Huda SN, Khatun F, Grantham-McGregor SM. Psychosocial stimulation improves the development of undernourished children in rural Bangladesh. *Journal of Nutrition*. 2006; 136: 2645–2652.

15. Aboud FE, Akhter S. A cluster-randomized evaluation of a responsive stimulation and feeding intervention in bangladesh. *Pediatrics*. 2011; 127: 1191–1197.

16. Nahar B, Hamadani JD, Ahmed T, et al. Effects of psychosocial stimulation on growth and development of severely malnourished children in a nutrition unit in Bangladesh. *European Journal of Clinical Nutrition*. 2009; 63: 725–731.

17. Nahar B, Hossain MI, Hamadani JD, et al. Effects of a community-based approach of food and psychosocial stimulation on growth and development of severely malnourished children in Bangladesh: a randomised trial. *European Journal of Clinical Nutrition*. 2012; 66: 701–709.

18. Powell C, Baker-Henningham H. Feasibility of integrating early stimulation into primary care for undernourished Jamaican children: cluster randomised controlled trial. *BMJ.* 2004; doi:10.1136/bmj.38132.503472. & C (published 24 June 2004)

19. Walker S, Chang S, Powell C, Grantham-McGregor SM. Psychosocial Intervention Improves the Development of Term Low-Birth-Weight Infants. *Journal of Nutrition*. 2004; 134:1417–1423.

20. Meeks Gardner JM, Powell CA, Baker-Henningham H, Walker SP, Cole TJ, Grantham-McGregor SM. Zinc supplementation and psychosocial stimulation: effects on the development of undernourished Jamaican children. *American Journal Clinical Nutrition*. 2005; 82: 399–405.

21. Vazir S, Engle P, Balakrishna N, Husain N. Cluster-randomized trial on complementary and responsive feeding education to caregivers found improved dietary intake, growth and development among rural Indian toddlers. *Maternal and Child Nutrition*. 2013; 9: 99–117.

22. Lozoff B, Smith JB, Clark KM, Perales CM, Rivera F, Castillo M. Home intervention improves cognitive and social-emotional scores in iron-deficient anemic infants. *Pediatrics*. 2010; 126: 884–894.

23. Aboud FE, Yousafzai AK. Global health and development in early childhood. *Annual Review Psychology*. 2015; 66: 433–57.

24. Tofail F, Hamadani JD, Mehrin F, Ridout DA, Huda SN, Grantham-McGregor SM. Psychosocial stimulation benefits development in nonanemic children but not in anemic, iron-deficient Children. *Journal of Nutrition*. 2013; 143: 885–893.

25. Walker SP, Grantham-Mcgregor SM, Powell CA, Chang SM. Effects of growth restriction in early childhood on growth, IQ, and cognition at age 11 to 12 years and the benefits of nutritional supplementation and psychosocial stimulation. *Journal Pediatriscs*. 2000; 137: 36–41.

26. Walker SP, Chang SM, Powell CA, Grantham-McGregor SM. Effects of early childhood psychosocial stimulation and nutritional supplementation on cognition and education in growth-stunted Jamaican children: prospective cohort study. *Lancet*. 2005; 366: 1804–1807.

27. Walker SP, Chang SM, Vera-Hernández M, Grantham-McGregor SM. Early childhood stimulation benefits adult competence and reduces violent behavior. *Pediatrics*. 2011; 127: 849–857.

28. Yousafzai AK, Aboud F. Review of implementation processes for integrated nutrition and psychosocial stimulation interventions. *Annals of the New York Academy of Sciences*. 2013; 1308: 33–45.

29. Albers CA, Grieve AJ. Test Review: Bayley, N. (2006). Bayley Scales of Infant and Toddler Development- Third Edition. San Antonio, TX: Harcourt Assessment. *Journal of Psychoeducational Assessment*. 2007; 25: 180–190.

30. Hogrefe. Griffiths Mental Development Scales - Revised : Birth to 2 years (GMDS 0-2). 2015; http://www.hogrefe.co.uk/gmds-0-2.html

31. Prado EL, Abubakar AA, Abbeddou S, Jimenez EY, Somé JW, Ouédraogo JB. Extending the Developmental Milestones Checklist for use in a different context in Sub-Saharan Africa. *Acta Paediatr*. 2014; 103: 447–454.

32. Gladstone M, Lancaster GA, Umar E, et al. The Malawi Developmental Assessment Tool (MDAT): the creation, validation, and reliability of a tool to assess child development in rural African settings. *PLoS Medicine*. 2010; 7: e1000273.

33. Abubakar A, Holding P, van Baar A, Newton CR van de Vijver FJ. Monitoring psychomotor development in a resource-limited setting: an evaluation of the Kilifi Developmental Inventory. *Annal Tropical Paediatrics*. 2008; 28: 217–226.

34. Kariger P, Frongillo EA, Engle P, Britto PM, Sywulka SM, Menon P. Indicators of family care for development for use in multicountry surveys. *J Health Popul Nutr*. 2012; 30: 472–486.

35. Caldwell BM, Bradley RH. (2003) Home Observation for Measurement of the Environment: Administration Manual: Tempe, AZ: Family & Human Dynamics Research Institute, Arizona State University.

36. Yousafzai AK, Rasheed MA, Bhutta ZA. Annual Research Review: Improved nutrition-pathway to resilience. *Journal of Child Psychology and Psychiatry*. 2013; 54: 367–77.

37. Grantham-McGregor SM, Fernald LC, Kagawa RM, Walker S. Effects of integrated child development and nutrition interventions on child development and nutritional status. *Annals of the New York Academy of Sciences*. 2014; 1308: 11–32.

38. Bentley ME, Johnson SL, Wasser H, et al. Formative research methods for designing culturally appropriate, integrated child nutrition and development interventions: an overview. *Annals of the New York Academy of Sciences*. 2014; 1308: 54–67.

39. Mukerjee K. Early Childhood Care and Development (ECCD) in two communities in Tijuana, México: maternal and household characteristics and pediatric development [master's thesis]. Seattle: University of Washington; 2013.

40. Visión Mundial. Informe de síntesis: Evaluación del impacto del programa CPDI en las práctiacs familiares saludables para el cuidado y desaroollo de la primera infancia en colonias maginales de Tijuana, México. 18 de Septiembre del 2013.

41. Harris AD, Lautenbach E, Perencevich E. A systematic review of quasi-experimental study designs in the fields of infection control and antibiotic resistance. *Clinical Infectious Diseases.* 2005; 41: 77–82.

42. Harris AD, Bradham DD, Baumgarten M, Zuckerman IH, Fink JC, Perencevich EN. The use and interpretation of quasi-experimental studies in infectious diseases. *Clinical Infectious Diseases*. 2004; 38: 1586–1592.

43. Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one-group-per-condition-design: can any of the alternatives overcome the problems inherent in this design? *Evaluation Review*. 2001; 25: 440–453.

44. Ramírez, MA. Evaluación cualitativa del programa cuidado y desarollo de la primear infancia (CDPI) de visión mundial aplicado en el área desarrollo de projecto cañón del sainz, Tijuana, México: Reporte tecnico. Revisado 20 de marzo de 2013.

45. WHO Multicenter Growth Reference Study Group. WHO Motor Develoment Study: Windows of achievement for six gross motor milestones. Acta Paediatrica Supplement 2006;450:86-95

46. Radloff L. The CES-D scale a self-report depression scale for research in the general population. *Applied Psychological Measurement*. 1977; 1: 385–401.

47. Vázquez FL, Blanco V, López M. An adaptation of the Center for Epidemiologic Studies Depression Scale for use in non-psychiatric spanish populations. *Psychiatry Research*. 2007; 149: 247–52.

48. UNICEF. Mutiple Indicator Cluster Survery (MICS). UNICEF Web site. 2015. http://mics.unicef.org/ Accessed April 2015.

49. Christie J, Halloran PO, Stevenson M. Planning a Cluster Randomized Controlled Trial: Methodological Issues. *Nursing Research*. 2009; 58: 128–134.

50. Campbell MK, Piaggio G, Elbourne DR. Research Methods & Reporting: Consort 2010 statement, extension to cluster randomised trials. *BMJ*. 2012; 345: 1–21.

51. Walker SP, Wachs TD, Gardner JM, et al. Child development: risk factors for adverse outcomes in developing countries. *Lancet*. 2007; 369: 145–157.

52. Gertler P, Heckman J, Pinto R, et al. Childhood psychosocial stimulation in Jamaica enhances later wage earnings. *Journal of Pediatrics*. 2014; 16: 1070.

53. Wachs T, Black M, Engle P. Maternal Depression: A Global Threat to Children's Health, Development, and Behavior and to Human Rights. *Child Dev Perspect*. 2009; 3.

54. Surkan PJ, Kennedy CE, Hurley KM, Black MM. Maternal depression and early childhood growth in developing countries: systematic review and meta-analysis. *Bull World Health Organ*. 2011; 89: 608–615.

55. Sohr-Preston SL, Scaramella LV. Implications of timing of maternal depressive symptoms for early cognitive and language development. *Clinical Child and Family Psychology Review*. 2006; 9: 65–83.

Table 1. Definitions of covariates mentioned in thesis text or listed in following tables.	
Covariates [1,2]	Definition
Final Analysis[3]	
CHILD DEVELOPMENT	
Milestone attainment	Passed all milestones assessed vs. not on age specific questionnaire (dichotomous)
CHILD STIMULATION ACTIVITIES *For all stimulation activities, caregivers were asked if in general they or another family member engaged in the following stimulation activities with their children. No time frame or duration of stimulation was specified.*	
Play	Playing with child vs. not (dichotomous)
Listening to music	Listening to music with child vs. not (dichotomous)
Reading	Reading with child vs. not (dichotomous)
Singing	Singing with child vs. not (dichotomous)
Speaking	Speaking with child vs. not (dichotomous)
Massage	Massage/body movements with child vs. not (dichotomous)
CHILD STIMULATION TIME	
Stimulation time	Total amount of stimulation time undertaken with child by all family members during the previous day and night divided into three categories: no stimulation, ≤ 2 hours, and > 2 hours (categorical)
CHILD CHARACTERISTICS	
Iron-rich foods	Child was fed meat, fish, chicken or legumes in day preceding interview (dichotomous)
Dietary diversity	< 3 vs. ≥ 4 food groups from list of 7 possibilities (dichotomous)
Child age	Child age in months (continuous)
MATERNAL CHARACTERISTICS	
Education, incomplete primary school	Incomplete primary school vs. all other education categories (dichotomous)
Education, categorical	5 categories: incomplete primary, complete primary, incomplete secondary, complete secondary preparatory/university (categorical)
Employment outside the home	Maternal employment outside of home vs. not (dichotomous)
Employment as a maquiladora	Mother employed as a maquiladora vs. not (dichotomous)
Raw depression score	Depression score (possible range 0-60), adjusted for missing values (continuous)
Depression, categorical	Depression score in 3 categories: none (<16), moderate (16 – 23.99), severe (>24) (categorical)
Single parent household	Single parent household vs. not (dichotomous)
STUDY DESIGN	
Exposure to the intervention	Participated in one or more WV session(s) vs. not (dichotomous)
Time[5]	Endline vs. baseline data (dichotomous)
Arm[5]	Intervention vs. comparison (dichotomous)
Questionnaire	Questionnaire administered to household was designed for children of ages 6 – 11.99 months vs. ≥12 months (dichotomous)
Preliminary Analysis[4]	
CHILD DEVELOPMENT	
Language attainment	Number of words spoken by child (continuous, 12 – 35.99 mo. [available for ≥ 12 mo. only])
CHILD STIMULATION ACTIVITES AND TIME	
Diversity of activities	All stimulation activities combined into one countable diversity measure (scaled 0-6) (categorical)
Stimulation time	Expanded categorical version of stimulation time variable, here measured in multiples of 15 minutes (scale: none, 15 min, 45 min, 90 min, 150 min) (categorical)

1. All variables described data derived from maternal self-report unless otherwise noted
2. All variables described pertain to children in the 6 – 23.99 month age group, unless otherwise noted
3. The subsection entitled "Final analysis" includes all variables considered in the final multivariate or bivariate analyses
4. The subsection entitled "Preliminary analysis" lists variables used in the preliminary analysis phases only
5. Bivariate and multivariate analyses were adjusted for a time/arm interaction term

Table 2. Maternal, child, and household characteristics for participants in comparison and intervention groups, at baseline and endline.[1,2]

Covariates (Overall # of missing values)	Comparison		Intervention		
	Baseline N = 120	Endline N = 36	Baseline N = 109	Endline, unexposed N = 96	Endline, exposed[3] N= 27
CHILD DEVELOPMENT					
Pass all (4)	64 (53.3%)	17 (48.6%)	52 (47.7%)	56 (59.6%)	17 (63%)
PSYCHOSOCIAL STIMULATION					
Play (24)	80 (66.7%)	20 (76.9%)	63 (59.4%)	53 (61.6%)	17 (65.4%)
Listening to music (31)	13 (10.8%)	9 (43.9%)	23 (21.9%)	37 (43%)	6 (24%)
Reading (39)	4 (3.4%)	9 (45%)	4 (3.9%)	16 (19.3%)	4 (16.7%)
Singing (25)	18 (15.3%)	21 (77.8%)	26 (25%)	48 (53.9%)	14 (56%)
Speaking (22)	51 (43.2%)	20 (69%)	28 (26.2%)	49 (55.1%)	11 (42.3%)
Massage (37)	47 (40.2%)	15 (71.4%)	38 (36.5%)	49 (58.3%)	13 (52%)
Total stimulation time >2hrs (27)	8 (7.1%)	9 (25%)	15 (15.3%)	29 (31.5%)	8 (32%)
CHILD CHARACTERISTICS					
Sex, female (8)	49 (42.2%)	18 (50%)	50 (46.7%)	55 (58.5%)	18 (66.7%)
Age in months (0)	14.9 (10.5-19.8)	13.8 (11.6-18.5)	14.6 (9.7-20.1)	15.9 (11.6-19.4)	14.9 (11.2-19.2)
Iron-rich foods (12)	62 (51.7%)	29 (96.7%)	50 (45.9%)	56 (61.5%)	12 (46.2%)
Dietary diversity (4)	35 (29.2%)	15 (44.1%)	32 (29.4%)	29 (30.9%)	6 (22.2%)
MATERNAL CHARACTERISITICS					
Education, incomplete primary school vs. all other categories (2)	14 (11.9%)	1 (2.8%)	14 (12.8%)	6 (6.3%)	2 (7.4%)
Employment outside the home (1)	39 (32.5%)	10 (27.8%)	35 (32.1%)	27 (28.4%)	6 (22.2%)
Employment as a maquiladora (2)	18 (15%)	6 (16.7%)	15 (13.8%)	18 (19.2%)	3 (11.1%)
Maternal depression (12)					
None	79 (67%)	21 (60%)	62 (57.4%)	55 (59.8%)	10 (38.5%)
Moderate	20 (17%)	13 (37.1%)	29 (26.9%)	22 (23.9%)	10 (38.5%)
Severe	19 (16.1%)	1 (2.9%)	17 (16.7%	15 (16.3%)	6 (23.1%)
Raw depression score (7)	12 (7-18)	12.6 (9-18)	13 (6-20)	13 (9-19)	18.5 (10.5-22)
Age at delivery, (0)					
<20 years	31 (25.8%)	3 (8.3%)	30 (27.5%)	25 (26%)	8 (29.6%)
20-34 years	79 (65.8%)	29 (80.6%)	71 (65.1%)	63 (65.6%)	14 (51.9%)
>34 years	10 (8.3%)	4 (11.1%)	8 (7.3%)	8 (8.3%)	5 (18.5%)
Single parent household (0)	16 (13.3%)	7 (19.4%)	16 (14.7%)	23 (24%)	5 (18.5%)

1. Not adjusted for child age or design variables
2. Data presented in the following format: dichotomous or categorical variables n (%), and continuous variables median (IQR)
3. Endline, exposed: Participated in ≥ 1 parental education session in context of World Vision intervention

Table 3. Combined bivariate analysis (planned analyses) and stratified bivariate analysis (exploratory analyses).[1 2 3]

Covariates	Combined analysis (all participants) N = 396	Stratified analyses			
		Comparison		Intervention	
		Baseline N = 120	Endline N = 36	Baseline N = 109	Endline N = 131
STIMULATION TIME					
No stimulation	reference	reference	reference	reference	reference
Some stimulation, ≤2hrs	1.25 (0.58, 2.71)	0.71 (0.23, 2.17)	Undefined	2.40 (0.62, 9.33)	2.77 (0.27, 28.37)
>2hrs of stimulation	2.18 (0.84, 5.67)	1.58 (0.23, 10.63)	Undefined	3.76 (0.54, 26.05)	3.43 (0.31, 37.62)
CHILD STIMULATION ACTIVITIES					
Play	1.84 (1.15, 2.93)	3.78 (1.59 8.95)	Undefined	1.52 (0.64, 3.61)	0.72 (0.30, 1.73)
Reading	1.04 (0.47, 2.31)	0.45 (0.06, 3.54)	1.89 (0.28, 12.98)	1.28 (0.16, 10.35)	0.85 (0.28, 2.57)
Listening to music	1.73 (0.997, 3.01)	1.71 (0.47, 6.24)	2.39 (0.33, 17.43)	2.14 (0.77, 5.94)	1.38 (0.60, 3.19)
Singing	1.19 (0.71, 1.99)	0.34 (0.11, 1.06)	Undefined	1.15 (0.44, 2.99)	1.95 (0.86, 4.42)
Speaking	1.32 (0.83, 2.09)	1.17 (0.54, 2.56)	2.19, (0.29, 16.32)	1.2 (0.47, 3.08)	1.45 (0.63, 3.33)
Massage	0.84 (0.52, 1.35)	0.57 (0.25, 1.29)	Undefined	0.45 (0.17, 1.17)	1.35 (0.60, 3.06)
CHILD CHARACTERISTICS					
Sex, female	1.22 (0.79, 1.88)	1.29 (0.59, 2.81)	0.30 (0.07, 1.35)	1.31 (0.57, 3.02)	1.59 (0.72, 3.55)
Iron-rich foods	1.48 (0.93, 2.36)	1.82 (0.83, 4.00)	Undefined	1.60 (0.69, 3.70)	1.19 (0.51, 2.77)
Dietary diversity	1.72 (1.04, 2.83)	2.94 (1.13, 7.66)	0.59 (0.11, 3.02)	1.21 (0.49, 3.03)	2.17, (0.83, 5.66)
MATERNAL CHARACTERISTICS					
Maternal depression score	0.97 (0.94, 0.99)	0.97 (0.93, 1.02)	0.92 (0.81, 1.04)	0.97 (0.93, 1.02)	0.96 (0.92, 1.00)
Maternal employment outside the home	1.18 (0.74, 1.89)	1.02 (0.45, 2.32)	1.34 (0.23, 7.77)	1.85 (0.76, 4.51)	0.89 (0.37, 2.13)
Maternal employment at maquiladora	1.15 (0.64, 2.09)	1.44 (0.47, 4.40)	1.09 (0.16, 7.38)	1.17 (0.35, 3.93)	1.01 (0.37, 2.76)
Age at delivery,					
<20 years	1.20 (0.72, 2.01)	2.20 (0.88, 5.47)	0.63 (0.04, 9.28)	1.11 (0.43, 2.89)	0.79 (0.31, 2.00)
20-34 years	reference	reference	reference	reference	reference
>34 years	0.87 (0.41, 1.87)	1.59 (0.39, 6.56)	Undefined	1.64 (0.30, 9.06)	0.78 (0.22, 2.72)
Single parent household	0.90 (0.51, 1.59)	0.57 (0.18, 1.79)	0.50 (0.08, 3.24)	0.84 (0.26, 2.70)	1.53 (0.60, 3.91)
Maternal education,					
incomplete primary	0.64 (0.29, 1.43)	0.90 (0.23, 3.51)	undefined	0.25 (0.06, 1.10)	6.15 (0.96, 39.38)
complete primary	0.94 (0.54, 1.66)	1.41 (0.49, 4.07)	0.06 (0.00, 0.72)	0.91 (0.30, 2.71)	1.49 (0.53, 4.16)
incomplete secondary	0.63 (0.30, 1.32)	0.60 (0.17, 2.16)	0.11 (0.01, 1.70)	0.40 (0.08, 2.12)	1.50 (0.37, 6.14)
complete secondary	reference	reference	reference	reference	reference
preparatory or university	0.99 (0.52, 1.88)	1.56 (0.45, 5.39)	0.14 (0.01, 2.62)	1.18 (0.36, 3.90)	0.80 (0.25, 2.52)
OTHER COVARIATES					
Child age in months	1.20 (1.12, 1.30)	1.11 (0.98, 1.27)	1.17 (0.94, 1.46)	1.25 (1.09, 1.43)	1.29 (1.11, 1.50)
Questionnaire	0.70 (0.33, 1.48)	1.42 (0.37, 5.40)	0.87 (0.09, 8.40)	0.38 (0.09 1.67)	0.41 (0.11, 1.60)
Exposure to intervention	1.17 (0.46, 2.99)	N/A	N/A	N/A	1.22 (0.45, 3.29)

1. With the exception of section headed "Other covariates," all bivariate models investigate the individual association of each covariate of interest with milestone attainment, and are adjusted for child age and design variables (questionnaire, arm/time interaction term)
2. All study participants between 6 – 23.99 months of the appropriate time/arm who had information regarding each specific covariate were included in each model (N varied by covariate, data not shown)
3. Data are presented as odds ratio for association with milestone attainment (95% CI)

Table 4. Model 1a: Intended "final" multivariate model (preplanned analyses) with difference-of-differences analyses.[1 2 3]

	N	OR	95% CI	p value
Play	351	1.72	1.06, 2.81	0.029
Listening to music		1.58	0.88, 2.84	0.124
Maternal depression		0.96	0.94, 0.99	0.006
Iron-rich foods		1.38	0.84, 2.25	0.204
Change in comparison community[4]		0.26	0.08, 0.90	0.033
Change in intervention community[5]		1.83	0.99, 3.35	0.052
Impact attributable to WV program ("difference of differences")[6]		6.94	1.79, 26.88	0.005

1. Primary outcome: milestone attainment
2. All study participants between 6 – 23.99 months with responses to covariates in this model were included
3. The model is adjusted for child age, questionnaire, and the time/arm interaction
4. Odds ratio for change in developmental outcomes from baseline to endline in comparison community alone (derived from time/arm interaction term)
5. Odds ratio for change in developmental outcomes from baseline to endline in intervention community alone (derived from time/arm interaction term)
6. Odds ratio for difference in change in developmental outcomes from baseline to endline in intervention vs. comparison community (derived from time/arm interaction term)

Table 5. A comparison of model 1a (intended "final" multivariate model) with the series of alternative multivariate models (exploratory analyses).[1 2 3]

	Model 1			Model 2			Model 3[4]	Model 4[4]
	All participants Model 1a[4] N = 351	*Comparison arm only* Model 1c[5] N = 137	*Intervention arm only* *Model 1i*[5] N = 214	*All participants* Model 2a[4] N = 351	*Comparison arm only* Model 2c[5] N = 141	*Intervention arm only* *Model 2i*[5] N = 210	N = 375	N = 390
Play	1.72 (1.06, 2.81)	4.34 (1.82, 10.34)	1.05 (0.57, 1.95)					
Listening to music	1.58 (0.88, 2.84)	1.46 (0.45, 4.73)	1.62 (0.81, 3.23)					
Maternal depression	0.964 (0.94, 0.99)	0.97 (0.927, 1.01)	0.96 (0.93, 0.99)	0.969 (0.94, 0.995)	0.966 (0.92, 1.01)	0.969 (0.94, 1.00)	0.966 (0.94, 0.99)	
Iron-rich foods	1.38 (0.84, 2.25)	1.59 (0.68, 3.67)	1.31 (0.70, 2.46)	1.55 (0.94, 2.54)	1.58 (0.69, 3.62)	1.63 (0.865, 3.09)	1.54 (0.96, 2.48)	
Stimulation time: None				reference	reference	reference		
Stimulation time: ≤ 2 hours				1.30 (0.59, 2.87)	0.70 (0.23, 2.19)	2.51 (0.76, 8.27)		
Stimulation time: > 2 hours				2.06 (0.77, 5.53)	2.69 (0.51, 14.19)	2.79 (0.73, 10.70)		
Change in comparison community[6]	0.26 (0.08, 0.90)	0.19 (0.49, 0.74)		0.83 (0.17, 1.11)	0.36 (0.13, 0.99)		0.55 (0.22, 1.36)	0.93 (0.41, 2.11)
Change in intervention community[7]	1.83 (0.99, 3.35)		1.91 (1.03, 3.54)	1.53 (0.82, 2.85)		1.58 (0.83, 3.01)	2.03 (1.13, 3.66)	1.92 (1.09, 3.36)
Impact attributable to WV program ("difference-of-differences)[8]	6.94 (1.79, 26.88)			3.48 (1.17, 10.30)			3.70 (1.27, 10.78)	2.05 (0.77, 5.48)

1. Primary outcome: milestone attainment
2. All study participants between 6 – 23.99 months with responses to covariates in this model were included
3. Data are presented as odds ratio for association with milestone attainment (95% CI)
4. The model is adjusted for child age, questionnaire, and the time/arm interaction
5. Analysis stratified by study arm, this model is adjusted for child age and questionnaire
6. Odds ratio for change in developmental outcomes from baseline to endline in comparison community alone (derived from time/arm interaction term)
7. Odds ratio for change in developmental outcomes from baseline to endline in intervention community alone (derived from time/arm interaction term)
8. Odds ratio for difference in change in developmental outcomes from baseline to endline in intervention vs. comparison community (derived from time/arm interaction term)

Table 6. Comparing model 2 with alternate model 2 (unadjusted for baseline vs. endline).				
	Comparison		Intervention	
	Includes time Model 2c[1 2 3] N = 141	Excludes time Alternate 2c[1 2 3] N = 141	Includes time Model 2i[1 2 3] N = 210	Excludes time Alternate 2i[1 2 3] N = 210
Maternal depression	0.97 (0.92, 1.01)	0.97 (0.93, 1.01)	0.97 (0.94, 1.00)	0.97 (0.94, 1.00)
Iron-rich foods	1.58 (0.69, 3.62)	1.16 (0.54, 2.51)	1.63 (0.865, 3.09)	1.69 (0.54, 2.51)
Stimulation time: None	reference	reference	reference	reference
Stimulation time: ≤ 2 hours	0.70 (0.23, 2.19)	0.61 (0.20, 1.84)	2.51 (0.76, 8.27)	2.90 (0.91, 9.30)
Stimulation time: > 2 hours	2.69 (0.51, 14.19)	1.67 (0.34, 8.17)	2.79 (0.73, 10.70)	3.48 (0.96, 12.69)
Change in comparison community[4]	0.36 (0.13, 0.99)			
Change in intervention community[5]			1.58 (0.83, 3.01)	

1. Primary outcome: milestone attainment
2. All study participants between 6 – 23.99 months with responses to covariates in this model were included
3. The model is adjusted for child age and questionnaire
4. Odds ratio for change in developmental outcomes from baseline to endline in comparison community alone (derived from time/arm interaction term)
5. Odds ratio for change in developmental outcomes from baseline to endline in intervention community alone (derived from time/arm interaction term)