Journal of Development Economics

Registered Report Stage 1: Proposal

Evaluating the Effects of an Early Literacy Intervention

Authors: Pamela Jakiela (Center for Global Development, BREAD, IZA); Owen Ozier (World Bank Development Research Group, BREAD, IZA); Lia C. H. Fernald (University of California at Berkeley School of Public Health); Heather A. Knauer (University of Michigan School of Social Work)

Date of latest draft: January 10, 2020

Keywords: human capital, dialogic reading, emergent literacy, child development, storybooks, mother tongue, bilingual, multilingual

JEL codes: I25, O15

Study pre-registration: AEARCTR-0004425 (`https://doi.org/10.1257/rct.4425-1.0`)

Abstract: We conduct a cluster-randomized evaluation of an early literacy intervention that provided Kenyan parents with illustrated children's storybooks and modified dialogic reading training. Rural communities were randomly assigned to treatment or control. Within treatment communities, households were further randomized to receive children's storybooks in either Luo (the mother tongue of all children in the sample) or English (a national language, and the primary language of instruction in grade 4 of primary school and beyond). We estimate the impacts of treatment on children's vocabulary and literacy skills. Our design also allows us to document household responses to the intervention including behavioral responses by parents and older siblings and overall impacts on parental time investments in children.

EMERGE Project Timeline

2014:	Development of the intervention
2015:	Pilot study (individually-randomized in 9 communities)
2016:	Analysis of data from pilot study
	Construction of sample frame for present (cluster-randomized) study
2017:	Child censuses in study communities
	Baseline data collection in 73 communities (June–December, 2017)
2018:	Intervention delivered in 36 of 73 communities (February–March, 2018)
2019:	Endline survey (July 2019–January 2020 or beyond)
2020:	Data analysis

1 Introduction

Almost half of all children in low- and middle-income countries are at risk of failing to meet their developmental potential, primarily because of a lack of nutrition and early childhood stimulation (Black et al. 2017, Grantham-McGregor et al. 2007). Disparities in human capital begin early in life, often growing larger over time (Paxson and Schady 2007, Galasso, Weber, and Fernald 2019, Reynolds et al. 2017). The lack of adequate investment in early childhood leads to worse school performance, lower human capital accumulation, and reduced productivity in adulthood — creating a poverty trap for both individuals and societies (Behrman et al. 2014; Hanushek and Woessman 2012).

Early interventions can have lasting impacts, with potentially compounding effects over time (Carneiro and Heckman 2003, Almond and Currie 2011, Alderman et al. 2017). However, the best-known examples of successful early interventions in low- and middle-income countries (LMICs) are expensive (Gertler et al. 2014). In contrast, interventions that are feasible at scale often fail to reach the most vulnerable, or have short-run impacts that fade out as children age (Andrew et al. 2018; Attanasio et al. 2014; Dillon et al. 2017; Martinez, Naudeau, and Pereira 2012, 2017; Özler et al. 2018; Wolf et al. 2019). Thus, it is clear that early intervention is crucial, but there is limited evidence that low-cost, scaleable programs can generate lasting effects.

Children's books are a simple yet remarkable technology that wealthy parents take for granted. Across Sub-Saharan Africa, only 3 percent of children live in households that own at least three children's books; in higher-income countries, nearly all children do (UNICEF 2017).[1] Books nudge parents to engage with their children in the sort of back-and-forth conversations that are needed to spark self-sustaining growth in vocabulary, literacy, and other cognitive skills (Ninio 1983, Wasik and Hindman 2015). In the United States, distribution of children's books to low-income families has been shown to improve child vocabulary (High et al. 2000; Mendelsohn et al. 2001; Weitzman et al. 2004). For LMICs, the World Bank has identified universal literacy among children as a key policy goal (World Bank 2019). Though many literacy programs exist in LMICs, they typically target school-age children — rather than younger children who are building vocabulary and other

[1] See Table 12 of UNICEF (2017) for country-level statistics on the fraction of children under five years old who have at least three children's books: for example, 92 percent of children in Belarus have at least three books; that figure is 6 percent in Ghana, 3 percent in Mozambique, and only 1 percent in Rwanda.

pre-literacy skills, but not yet learning to read (Dowd et al. 2013; Piper et al. 2018). In Africa in particular, many parents are unaware of the importance of reading with children who have not yet reached school age, and interactive conversations between adults and young children are sometimes discouraged (Lancy 2015; Weber, Fernald, and Diop 2017; Jukes et al. 2018).

We are conducting a randomized trial examining the impacts of a low-cost parent education and storybook distribution program on young children in rural Kenya. Encouraging Multilingual Early Reading as the Groundwork for Education (EMERGE) is a cluster-randomized evaluation of a program that provides Kenyan parents with illustrated children's storybooks and modified dialogic reading training. The core question is whether this program has impacts on children's human capital — specifically, vocabulary and early literacy skills — one and a half years after the intervention took place. We test two variants of the intervention: in one, storybooks are printed in Luo (the local mother tongue), which might prove advantageous for young children (and caregivers) who only speak Luo; in the other variant, storybooks are printed in English, which might prove advantageous in terms of school readiness.

Before launching the present cluster-randomized trial, we documented the causal impacts of the EMERGE intervention on book-sharing practices through an individually-randomized pilot study. In 2014, we collaborated with a Kenyan publisher and dialogic reading experts to develop storybooks and a parent education program appropriate for this context. In 2015, we tested four variants of the EMERGE (storybooks plus dialogic reading) intervention in a pilot trial in nine communities in the same vicinity as the present study. That trial, documented in Knauer et al. (2019a), establishes the intervention's impacts on a five- to six-week timescale. Weeks after the intervention, nearly every intervention book was still in the (treatment) home that had received it; parents reported reading with their children at a much higher rate in treatment households than in control households; children in treated households knew the stories in the books; and treated children had richer vocabularies than their control group counterparts. These measures represent intermediate outcomes: they are the first steps in any reasonable theory of change linking the EMERGE intervention to improvements in children's human capital. The 2015 trial tested four different treatment intensities (books alone, books and parent education, and books plus more intensive assistance for parents), informing the present study by identifying the variant of the

EMERGE treatment that struck the best balance between cost and short-run impact.[2]

The present study measures the impacts of the EMERGE intervention approximately 18 months after treatment. In 2017, we conducted baseline surveys of 2,527 children aged 36 to 83 months in 2,013 rural households. We also surveyed each child's primary caregiver and collected data on all the other children then living in the household. After conducting the baseline survey, we randomly assigned treatment (storybooks plus dialogic reading training) at the community level. Within each treatment community, we also randomized the language of the storybooks across treatment households: storybooks were either in English (one of Kenya's national languages and the primary language of instruction in upper primary school) or Luo (the mother tongue of all children in our sample). Storybooks were produced for the EMERGE intervention and were identical in every respect except for the language that the story was printed in.

Whenever pilot interventions are scaled up, implementation fidelity is a potential concern. In the present study, members of the research team oversaw the implementation of the intervention and were able to monitor content delivery to ensure treatment fidelity. In addition, we conducted a midline survey in a small sample of households, again roughly five to six weeks after the intervention, to confirm that treatment led to the expected changes in reading behaviors. As in the pilot study, nearly every book was still in the treated homes six weeks after the intervention; parents reported increased reading frequency; and children knew the stories.

Our primary research question is whether the EMERGE intervention improved young children's vocabulary and early literacy skills. Through a two-level nested randomization, we also compare the impacts of English-language storybooks and mother tongue (specifically, Luo-language) storybooks. Language is an important but often overlooked issue in Kenya and in many developing country contexts. In rural Kenya, as in many other countries, children grow up speaking one of many locally spoken languages before learning official national languages in school. The most suitable language in which to present any early intervention is not obvious *ex ante*: books in a locally spoken language will be more readily familiar to children (if, for example, parents read the words printed in the books aloud), but availability of texts in an official national language may accelerate school readiness. To address this question, we randomize storybook language at

[2]Intervention costs were estimated at $28.27 per household in Knauer et al. 2019a.

the household level within treatment communities. This permits us to estimate treatment effects separately by the language of the children's books.

Ours is not the only study in economics to consider the importance of "mother tongue" language in early interventions, nor is it the only study of early literacy interventions in developing countries. For example, Kerwin and Thornton (2019) find that a mother-tongue literacy intervention in Uganda can have sizeable positive or negative effects, depending on details of outcome measurement and intervention intensity. In a similar vein, Lucas, McEwan, Ngware, and Oketch (2014) show that in Kenya and Uganda, the magnitude of the effect of literacy programs might depend on the language of instruction and the language of assessment. And while Piper et al. (2016, 2018) demonstrate a large effect of a mother-tongue intervention in Kenya, Chicoine (2019) offers a cautionary tale about the challenges of mother-tongue instruction in Ethiopia when such interventions may involve a change in alphabet.[3]

A critical feature of many early childhood interventions is that impacts on young children are mediated by changes in the behavior of older household members. Mothers, fathers, grandparents, and even siblings must be transformed into agents of change, so interventions can only work when these individuals change their parenting and caregiving practices. Though this design feature suggests that parent and sibling behaviors are critical mechanisms worth measuring, these pathways are rarely explored.[4] Moreover, limiting attention to impacts on young children may miss important benefits and costs of interventions: in a context where only half of mothers are literate, there may be positive spillovers on caregivers' literacy. On the other hand, time devoted to better parenting may reduce the time available for household tasks and leisure. To explore these issues, we measure child stimulation practices, time use, and literacy among both mothers and older siblings at endline. In doing so, we aim to provide a fuller picture of how intra-household responses to our intervention are optimized, so that any resulting understanding of costs and benefits is more comprehensive.

Registered reports and pre-analysis plans can be submitted at a range of project stages, conceivably as early as a grant proposal is funded. Early-stage reports must not only bear the

[3]The present study of course also contributes to the broader economics literature on the impacts and importance of early interventions (e.g., Almond and Currie 2011, Carneiro and Heckman 2003, Gertler, et al. 2014).

[4]Recent work by Das, Dercon, Habyarimana, Krishnan, Muralidharan, and Sundararaman (2013) is an exception, showing that an optimizing household's behavior can sometimes even counteract beneficial effects of interventions.

burden of anticipating state-contingent plans in the face of an exponentially expanding series of choices down the as-yet-unseen decision tree (Olken 2015), but must also describe how future implementation risks are managed. Because we are submitting our pre-analysis plan at a late stage of a project — and after many years of piloting — these risks are less relevant. As discussed above, development of the EMERGE project has been ongoing for five years, during which time we have been able to gather a tremendous amount of data documenting the validity of our measurement instruments; the fidelity of the intervention as implemented in the present, larger-scale, cluster-randomized trial; and the short-term impacts of the intervention on intermediate outcomes, both in the early pilot study and in the present cluster-randomized trial (e.g., Knauer et al. 2019a,b). This places the present report in a strong position with respect to many standard concerns. There is no risk that this intervention doesn't get implemented; no risk that it gets implemented differently than expected; and no risk that it doesn't translate into the short-term intermediate outcomes upon which our theory of change depends. Those risks have already been addressed. However, the critical, policy-relevant research question — whether a low-cost program combining storybooks with parent education leads to increases in children's human capital — still remains to be answered.[5]

To fully exploit the benefits offered by pre-analysis plans, we use this document to bind our hands in ways that increase the statistical power of our study, but would not be credible in the absence of a commitment technology. Specifically, we commit to the use of an explicit decision rule to determine which of two feasible specifications should be used in our endline analysis (in short, whether to include in our analysis a large sub-sample of children for whom no baseline data is available). We are not aware of an antecedent to its state-contingent nature. This innovation explores the potential value in reports such as this one: we maximize analytical possiblities, inasmuch as we can forsee them, while avoiding any p-hacking.[6]

In Section 2, we describe the EMERGE intervention in detail, and summarize the results of a short-term pilot study conducted prior to initiating this cluster-randomized trial. In Section 3,

[5]In relation to costing, note that we have access to detailed records of implementation costs associated with this intervention in both our previous pilot work and in the current larger-scale intervention. This information will be valuable in an examination of program cost-effectiveness, once we have estimated the program's impacts.

[6]Leaver, Ozier, Serneels, and Zeitlin (2020) use blinded endline data to choose among specifications on the basis of anticipated standard errors; in contrast, via this registered report, we commit to a procedure that we will use with the actual endline data, but we do so without yet having done analysis of endline data, blinded or otherwise.

we describe our sample, data collection, and treatment assignment procedures. Section 4 outlines our empirical strategy and hypotheses. Section 5 provides administrative details.

2 Research Design

2.1 Context

Literacy is important as a foundation for education, and through education, as a foundation for growth and development (World Bank 2019). Yet in Kenya—though it is one of the best-educated countries in Sub-Saharan Africa—less than a third of third-grade students can read at the second-grade level (Piper 2010) and only 34 percent of pre-school children are "on track" in terms of language and numeracy development (Kenya National Bureau of Statistics, 2013).

Our study takes place in rural areas of Kisumu County, a Luo-speaking region of western Kenya. Approximately 48 percent of the county's population lives below the local poverty line, and only 13 percent of adults completed secondary school (Commission on Revenue Allocation 2011). With over four million native speakers, Luo is Kenya's second-most-widely spoken language (Lewis et al. 2016). In rural areas of Kisumu County, 94 percent of the population speaks Luo as their mother tongue (as shown in data from the 2014 Demographic and Health Survey). High linguistic homogeneity made it feasible to implement a mother tongue storybook treatment without needing to translate both the books and the child assessments into multiple local languages.

Kenya's education policy stipulates that instruction should be given in the local mother tongue in the early years of primary school, but only 31 percent of young primary students in our study area are actually taught in Luo (Piper and Miksic 2011). Mother tongue instruction is ideal from the perspective of early learning (Ball 2010). It can also make it easier for parents to engage with their child's educational materials — since almost half of Kenyan mothers cannot read English at a second-grade level (Uwezo 2015). However, Kenyan parents often oppose mother tongue instruction because they believe it puts children at a disadvantage relative to those who do their schooling in English or Swahili (Trudell 2007, Jones 2012).

In our study area and elsewhere in rural Sub-Saharan Africa, many parents don't appreciate the importance of reading to preschool-aged children — whether in mother tongue or the official

language — and cultural norms may even discourage the kind of unstructured conversations that have been shown to spur language development in young children (LeVine et al. 2006; Lancy 2015; Weber, Fernald, and Diop 2017). Though literacy programs exist in many African countries (cf. Literacy Boost, Tusome, and PRIMR), most target children of primary school age, not during the pre-literacy period (Dowd et al., 2013; Piper et al., 2018; Piper et al., 2015). Very few young children even have access to age-appropriate reading materials at home: UNICEF's Multiple Indicator Cluster Surveys (MICS) show that only 4.4 percent of children's homes in this part of western Kenya had at least three children's books (Kenya National Bureau of Statistics, 2013). Though MICS data do not record storybook language, most children's storybooks available in Kenya are either in English or Swahili. Prior to our study, no Luo-language storybooks intended for preschool-aged children were available for sale anywhere in the Kisumu area.

Existing evidence suggests that early literacy materials should be made available in children's mother tongue. However, the absence of mother tongue storybooks from local bookstores combined with parents well-documented opposition to mother tongue instruction raised the possibility that households would prefer to receive early literacy materials in English or Swahili (Kenya's national languages). A second issue was how to design a parent education program that changed parents attitudes about the importance — and appropriateness — of reading to young children, and engaging them in informal, unstructured conversations during reading and at other times. As we discuss below, we were able to address these issues through extensive piloting, developing an intervention appropriate for the local context that leads to demonstrated changes in parent-child book-sharing behaviors.

2.2 Development of the Intervention

Our intervention combines two components: (i) locally-appropriate, illustrated **children's storybooks** in either English or Luo and (ii) a modified **dialogic reading training** that provided parents with guidance on how to engage and stimulate their young children through book-sharing. The intervention was developed by members of the research team in consultation with local stakeholders and dialogic reading experts. We provide a concise overview of each of the two intervention components below. A more detailed description of the intervention development process is

available in Knauer et al. (2019a).

2.2.1 Children's Storybooks

In 2014, we visited all of the major bookstores, markets, and grocery stores in the greater Kisumu area to assess the availability of children's storybooks in English and Luo. At that time, no Luo-language books intended for preschool-aged children were available at any location — though a limited set of English-language early readers (and relatively expensive imported books) were available in urban bookshops. To assess the potential demand for mother tongue storybooks, we began by translating and printing freely available content from the African Storybook Project (`https://www.africanstorybook.org/`). We distributed these and other locally-appropriate children's books to households in peri-urban Kisumu, then conducted semi-structured interviews and focus groups to understand parents' views of different types of storybooks.

Based on the feedback we received from parents, we partnered with Kenyan-owned Moran Publishers to adapt and translate six of their early readers (originally intended for primary school students). In our focus groups, many parents appreciated the detailed, colorful illustrations of African life depicted in the Moran books (see Figure 1). We adapted these books by modifying the English-language content to be appropriate for a parent reading together with younger children, and then produced parallel English and Luo editions. All storybooks contained embedded vocabulary words specific to the story (e.g. "umbrella") and questions intended to help parents start conversations about the plot with their young children.

2.2.2 Dialogic Reading

Shared reading is most effective when parents or teachers engage children in a dialogue — children build vocabulary skills more rapidly when they formulate their own questions about stories (Duursma, Augustyn and Zuckerman 2008). Dialogic reading is an approach to book-sharing that emphasizes children's active engagement, offering parents (or teachers) practical tools to encourage children to articulate their own questions and ideas about a story (Whitehurst et al. 1988; Zevenbergen and Whitehurst 2003). Dialogic reading programs have been shown to improve children's expressive vocabulary and emergent literacy skills in high-income countries (Mol et al.,

Figure 1: Sample Storybook: Illustration from Moran Publishers' *Market Day*©

2008). Most evaluations of dialogic reading programs in low- and middle-income countries have focused on classroom settings: for example, Oper, Ameer, and Aboud (2009); Elmonayer (2013); and Simsek and Erdogan (2015) find that classroom-based dialogic reading programs improved child vocabulary in Bangladesh, Egypt, and Turkey, respectively. In South Africa, a dialogic book-sharing intervention provided to mothers with children aged 14 to 18 months increased children's sustained attention and vocabulary (Murray et al. 2016; Vally et al. 2016).

To develop a modified dialogic reading training intervention appropriate to our context, child development specialists on the research team adapted materials from both the Oper, Ameer, and Aboud (2009) intervention in Bangladesh and the Vally et al. (2015) intervention in South Africa. The parent education program we developed presents the core elements of dialogic reading in a culturally-appropriate format that emphasizes the importance of engaging young children in book-centered conversations. The training materials specifically encourage caregivers with limited literacy: we emphasize the importance of engaging children through discussion of storybook content and illustrations — and the relative un-importance of reading the text word-for-word.

2.2.3 The EMERGE Intervention

The EMERGE intervention is a bundled treatment that combines locally-appropriate storybooks with modified dialogic reading training (both adapted from existing content by the research team, as described above). To further refine our intervention, we conducted a small, short-term pilot study in 2015. The pilot compared a control group (i) to the combination of storybooks and training, (ii) to a lighter-touch intervention that provided only storybooks without dialogic reading training, and (iii) to two more intensive interventions that also included additional booster trainings and home visits (Knauer et al. 2019a).[7] Variants of the EMERGE intervention that included both storybooks and training increased the frequency of parent-child book-sharing, improved the quality of book-sharing activities, and improved children's knowledge of vocabulary words embedded in the storybooks (measured six weeks after treatment). Delivering storybooks without offering parents our modified dialogic reading training increased the frequency of shared reading, but did not impact the quality of reading engagement or children's vocabulary.

In the present study, households assigned to treatment were invited to attend a modified dialogic reading training that was held in the local primary school or another central meeting place within the community. Storybooks were distributed to caregivers at the conclusion of the training. 88 percent of households assigned to treatment sent at least one adult to participate in the training. Whenever possible, we delivered storybooks to the homes of those caregivers who did not participate in the training — so a total of 97 percent of households assigned to treatment received either storybooks and modified dialogic reading training or storybooks alone. In short, take-up of the intervention is extremely high.

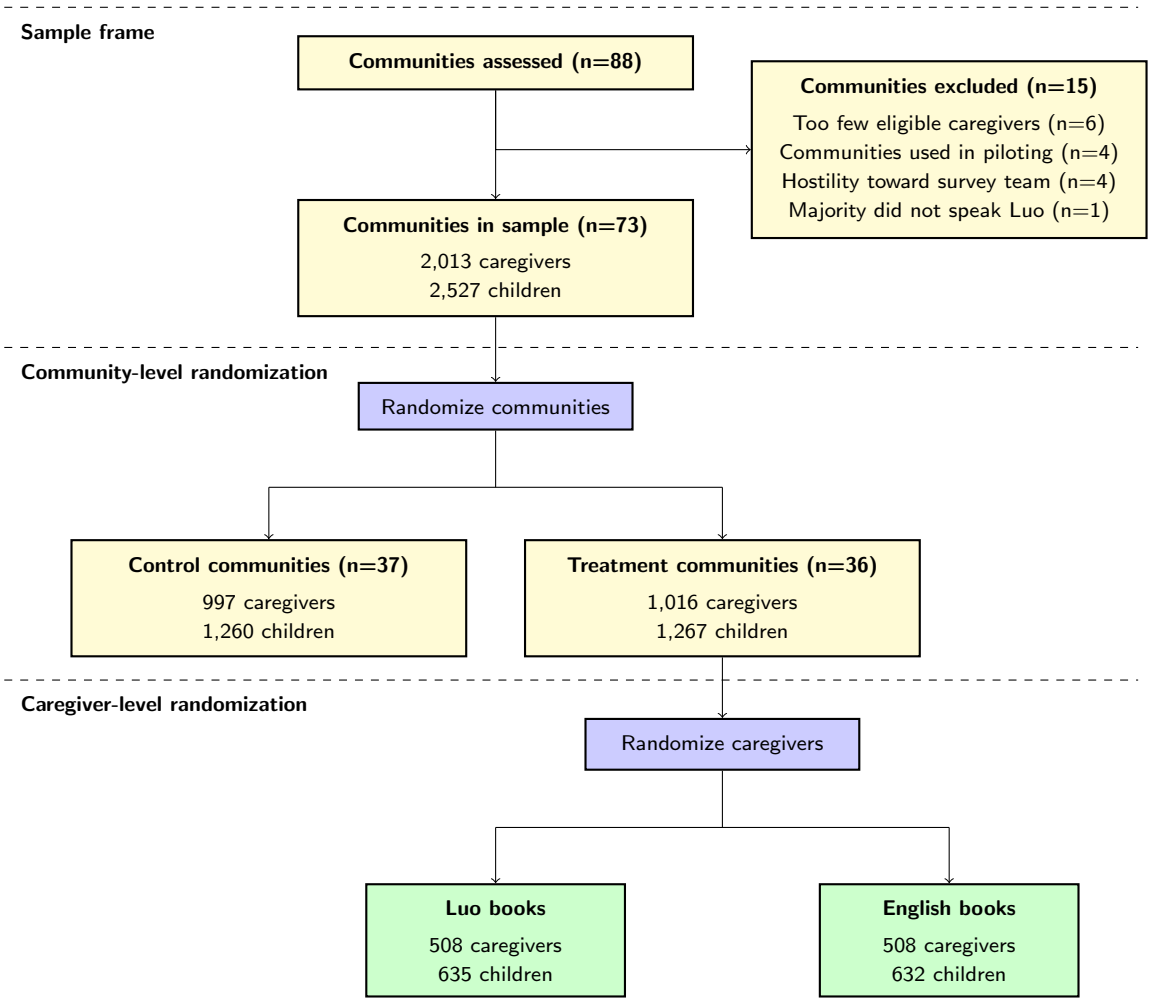
2.3 Identification Strategy

Our study is a cluster-randomized trial. Figure 2 summarizes the research design, a two-level nested randomization implemented in 73 rural communities in Nyando sub-county, Kenya. First,

[7]The pilot study was randomized at the individual level in nine rural communities. Children were assessed six weeks after treatment, limiting the potential for spillovers. (Parents assigned to the control arm in the pilot were invited to a second parent training session two months after treatment, i.e. two weeks *after* endline.) We found little evidence of control group contamination six weeks after treatment: only 13 percent of control households had any children's storybooks at endline, as compared to 97 percent of households assigned to (any) treatment; and children in the control group answered an average of 0.5 of 10 storybook comprehension questions correctly, as compared to an average of 3.7 correct responses in the treatment groups. Nonetheless, to fully eliminate the possibility of contamination, the present study was cluster-randomized at the community level, as discussed below.

communities were randomly assigned to either the treatment group or the control group. Within treatment communities, caregivers were further randomized into two treatment arms: Luo-language storybooks or English-language storybooks. All caregivers who completed the baseline in treatment villages were invited to attend the modified dialogic reading training. Books were distributed after the training according to the caregiver-level randomization. Books were also distributed to the homes of caregivers who chose not to attend the training.

Figure 2: Research Design



3 Experimental Procedures and Data

3.1 Sample

Our study area is a region of continuous settlement, and primary school catchment areas do not map to specific villages or other recognizable administrative units. After extensive pre-testing, we defined communities (for the purposes of our study) as the area within 750 meters of a primary school where parents were likely to send all of their children to that school; further from any school, one observes considerable variation in which (public) primary school children attend.

To construct a community-level (i.e. school-level) sample frame, we combined data from the 2007 Kenyan School Mapping Project with administrative data on the primary school leaving exam, the Kenya Certificate of Primary Education, to create a list of all public, coeducational day (i.e. not boarding) primary schools in Kisumu County, Kenya. After excluding those schools that were located in (linguistically heterogeneous) urban areas, we confirmed the location and status of all the rural schools through site visits, dropping schools from the sample frame that were no longer operational or too close to one another to plausibly avoid contamination across treatment arms (using a distance threshold 1.5 kilometers). We also excluded larger primary schools (with graduating class sizes of 20 or above) because we anticipated measuring classroom-level outcomes at endline (and wished to examine contexts were it was plausible to treat entire cohorts). 88 schools in three rural constituencies met our eligibility criteria; these schools constituted our initial community-level sample frame.

In 2017, we conducted censuses in each study community to construct a sample frame of children aged 36 to 83 months and their primary caregivers (typically mothers), mapping the location of all households within 750 meters of the primary school that met eligibility criteria (based on children's ages). We attempted censuses in all 88 communities, however activities were stopped in four locations because of community hostility. One additional community (on the border between Kisumu and Nandi counties) was dropped from the sample because a large proportion of caregivers were native Nandi speakers. A further six communities were dropped from the sample after the census because the number of eligible caregivers (with children between the ages of 36 and 83 months) was too small (10 or fewer). This left a sample of 77 communities,

four of which were used for pilot testing of our survey instruments and intervention protocols.

3.2 Baseline Data Collection

Between June 8 and December 21, 2017, the EMERGE field team conducted baseline surveys of 2,013 caregivers and 2,527 children aged 36 to 83 months.[8] In communities with fewer than 44 households with eligible-age children, we invited all eligible caregivers to participate in the study. In larger communities, we randomly sampled 44 caregivers for inclusion in the sample — though all caregivers of eligible children were invited to attend the dialogic reading training and receive storybooks. When a caregiver had one or two eligible-age children in her care, we conducted a survey with each child. In 46 (of 2,013) cases where a caregiver had more than two eligible children, we randomly sampled two for inclusion in the sample.

3.3 Baseline Characteristics

Baseline characteristics of the children in our sample are summarized in Table 1. The median household size is six. A typical household had an iron roof and a latrine within their compound at baseline, but did not have electricity, a cement floor, a bicycle, or a car.

86 percent of the children in the sample were cared for by their mother at baseline, and another 11 percent were looked after by their grandmother (typically because their mother was working full-time elsewhere; only 3 percent of children in the sample have a mother who was deceased at baseline). 50 percent of sample children have an illiterate primary caregiver. 95 percent of sample children have a mother whose native language is Luo. The median level of maternal education is eight years — i.e. complete primary school, but no secondary school.

50 percent of sample children are male. As expected, the median age is 60 months. The median height-for-age z-score was -0.48 at baseline, and 11 percent of children in the sample were stunted. 86 percent of children in the sample were enrolled in school at baseline.

[8] As we discuss further below, the political uncertainty surrounding Kenya's presidential election necessitated a two-month pause in surveying between July 28 and October 2, 2017.

3.4 Treatment Assignments

3.4.1 Community-Level Randomization

The community-level randomization was stratified as follows. First, we divided communities into two groups: those that completed baseline surveys prior to Kenya's presidential election (45 communities, surveyed in June and July of 2017) and those that completed baselines after the October re-run of the election (28 communities, surveyed in November and December of 2017).[9] Since election-related security concerns delayed some baseline surveys by several months, we stratified treatment assignments by baseline timing in order to reduce variation in delay between baseline and intervention. We further stratified communities by geography (the pre-election group was partitioned into northern and southern regions), community size, and school quality (as measured by the mean primary school leaving exam score in years prior to the intervention).

Preliminary analysis of the baseline data suggested that stratification was not sufficient to guarantee balance on key outcomes of interest. After stratifying by baseline timing, geography, community size, and school quality (as described above), key covariates such as school enrollment, child age, whether a child's mother was Luo, and the number of eligible children in the household were imbalanced (i.e. differences between treatment and control means were significant at the 10 percent level) up to 20 percent of the time in samples of 100 treatment assignments. We address this by using a re-randomization approach (Bruhn and McKenzie 2009, Athey and Imbens 2017). To assign communities to treatment and control groups, we generated one thousand stratified random assignments (using Stata 13.1), removing from consideration those that were imbalanced at the 10 percent level on any one of 12 key covariates (household size, mother's education, whether the mother is the primary caregiver, primary caregiver literacy, child gender, child age, child height-for-age z-score, school enrollment, expressive vocabulary, receptive vocabulary in Luo, parental stimulation of young children, and school quality). Of the first thousand random assignments we generated, 204 community-level treatment assignments met these balance criteria; we randomly chose one from this set. Following this procedure to generate additional alternative

[9]The first presidential election took place on August 8, 2017. That election was subsequently annulled by Kenya's Supreme Court on September 20. A new election was held on October 26. Because of security concerns during the period surrounding the election and the subsequent annulment, no baseline activities took place in August, September, or October.

treatment assignments will allow us to calculate randomization inference p-values anywhere that is desired as a robustness test. Since the space of eligible treatment assignments is quite large, and the re-randomization procedure does not constrain the space that severely, these p-values should be broadly comparable to those obtained through classical (asymptotic) inference; if anything, we expect the classical (asymptotic) inference to be conservative (Athey and Imbens 2017).

Table 2 reports baseline summary statistics separately for children in treatment and control communities (omitting those characteristics where balance was enforced). The treatment and control groups are generally well-balanced, though (as expected) there are a few differences. Children in control communities are slightly less likely to have an iron roof (99 percent in treatment vs. 96 percent in control), though the difference is small in magnitude and our aggregate asset index is balanced across treatment arms.[10] Children assigned to the control group were slightly more likely to have a Luo mother (94 percent in treatment vs. 96 percent in control), though again the difference is quite small in magnitude. Children assigned to treatment also perform slightly worse on the baseline math assessment, though pre-literacy skills (receptive vocabulary in English, familiarity with letters, and performance on familiar word reading tasks in English and Luo) and fine motor skills are similar in the treatment and control groups.

3.4.2 Caregiver-Level Randomization: Storybook Language

Within treatment communities, caregivers were randomly assigned to receive either English or Luo storybooks. Randomization was stratified by community. Because some treatment communities were quite small, scope for further stratification was limited. Instead, we once again employed re-randomization. Caregiver assignments were checked for balance in terms of household size, household assets, distance to the school, whether the primary caregiver was the child's father or grandmother, caregiver age, caregiver education, whether the caregiver was Luo, caregiver vocabulary in both Luo and English, caregiver numeracy, caregiver digit span, whether the child's mother was alive, whether the child's mother was Luo, the age of the child's mother, whether the child's preferred language was Luo, an index of preventive health investments (vaccinations, whether the child sleeps under a mosquito net, whether the child had been given vitamins or

[10]The asset index is the first principal component of a broad set of indicators for home quality and ownership of durable goods and livestock, following the procedures outlined by Filmer and Pritchett (2001).

deworming medication in the last six months), receptive vocabulary in English, familiarity with letters, brief early literacy assessments using familiar words in English and Luo, fine motor skills, and a short math skills assessment. In spite of the large number of balancing variables, 296 of one thousand initial random treatment assignments satisfied our balance criteria, allowing us to first use traditional (asymptotic) p-values, but also laying the groundwork for calculation of randomization inference p-values as robustness checks in our analysis, analogously to the community-level randomization and as discussed by Athey and Imbens (2017).[11]

3.4.3 Intervention Delivery

The baseline research team delivered the intervention to 36 randomly chosen communities in April and May of 2018. As discussed above, 88 percent of sample households that were assigned to treatment sent at least one adult household member to participate in training; we were able to deliver storybooks (without training) to the homes of a further 9 percent of treatment households.

3.4.4 Midline Data Collection

Several weeks after the intervention took place in 2018, we randomly sampled a small number of respondents for a follow-up survey to measure outcomes related to intervention fidelity (e.g. the presence of EMERGE storybooks in the home). We sampled half of the communities in two of the three geographic strata (22 communities), and within those, either sampled all households (for small-community strata) or a random half of households (for large-community strata). We then sampled one child to focus on per household (relevant when we had gathered baseline data on two children). Thus we sampled 394 caregivers and children, of whom we were able to conduct a midline survey for 379 (just over 96 percent).

The purpose of the midline data collection is **not** to preempt the endline: we did not gather any vocabulary or literacy measures, our any of our other focal endline outcomes. The purpose of the midline was to document intervention mechanisms and fidelity in the cluster-randomized study, permitting a direct comparison to our previous small-scale pilot, as discussed below.[12]

[11] Because of the large number of balancing variables, we omit the balance check table since balance is enforced for all variables of interest.

[12] Two risks, well-known in relation to books, are relevant to discuss here: that books could be unhelpful if their difficulty level is inappropriate, or that they could go entirely unused if parents perceive the books as being

Midline impacts are shown in Table 3. Being assigned to treatment increased the probability of having *any* children's books in the home by 88 percentage points, and increased the number of children's books found in the home by 4.59 (the number of books distributed was five). Children in the treatment group correctly answered an average of 4.57 more comprehension questions about the stories than did children in the comparison group (who correctly guessed an average of 0.4 questions out of 11).[13] Panel B of Table 3 shows that treatment increased the likelihood that someone read to a child in the past three days by 32 percentage points, comparable to the 22-26 percentage point increases that we observed in our pilot (see Knauer et al. 2019a, Table 2). Panel C of Table 3 shows midline impacts on reading behaviors. To measure this outcome, a survey team member observes the caregiver reading with a child, and codes the observed behaviors for twenty periods of 15 seconds each. As the table shows, we saw more interactive reading behaviors among caregivers assigned to the treatment group (almost all of whom received books and participated in training). We are thus confident not only that take-up was high (from administrative records), but also that intervention fidelity was good: midline impacts in the present study are comparable to those in our prior short-term study, in terms of books in the home; caregiver reports of reading; children's familiarity with stories; and observed reading behaviors.

3.4.5 Endline Data Collection

Adaptation of endline survey modules begin in early 2019. Since children had aged since the baseline survey and the endline survey also included older siblings, it was necessary to develop measures of vocabulary and early literacy appropriate for school-aged children. We also developed a new module measuring women's time use and secondary activities that captures the amount of time caregivers spend engaging with their young children.

The endline survey was launched on July 1, 2019. Data collection is expected to take approximately six months, concluding in early 2020. Survey teams are tracking 2,013 households containing 5,012 children aged 18 to 143 months at baseline: in addition to the 2,527 baseline children, the endline sample also includes 442 younger siblings (18 to 35 months at baseline) and

so valuable that children might not be allowed to interact with them at all (Glewwe, Kremer, and Moulin 2009; Sabarwal, Evans, and Marschak, 2014). Both of these risks are mitigated in this context: the effects seen in our midline data (and pilot study data), discussed above, investigate and refute both of these possible concerns.

[13]Sample story comprehension questions are shown in Appendix Figure A1.

2,043 older siblings (84 to 143 months at baseline).

4 Empirical Analysis

Our study is designed to answer two main research questions. First, does the combination of contextually-appropriate children's storybooks and dialogic reading training for primary caregivers lead to improvements in vocabulary and literacy? Second, are mother tongue storybooks more (or less) effective than storybooks in the national language?

4.1 Statistical Methods

To answer the first research question, we will estimate the OLS regression equation

$$Y_{ihv} = \alpha + \beta T_v + \gamma_s + X_{ihv} + \varepsilon_{ihv} \tag{1}$$

where Y_{ihv} is an outcome of interest (e.g. expressive vocabulary) for child i in household h in village (community) v, T_v is a treatment dummy equal to one if village v is assigned to treatment (storybooks plus dialogic reading training), γ_s is a vector of fixed effects for randomization strata, and X_{ihv} is a vector of baseline covariates.[14] Whenever possible, X_{ihv} will include the baseline value of the outcome variable. Since treatment was randomly assigned at the community level, standard errors must be clustered at the community level. Given our randomized design, $E[\hat{\beta}]$ is the average impact of being assigned to the EMERGE treatment, averaging over households receiving English books and Luo books. Equation 1 can also be used to estimate the impact of either the Luo-books or English-books treatment, as compared to the control group, by restricting the treatment sample to those randomly assigned to one of the two language sub-treatments (since assignment to treatment is randomized at both the community and the caregiver level).

To answer the second research question, we will restrict attention to treatment villages to estimate the OLS regression equation

$$Y_{ihv} = \eta + \delta L_{ihv} + \lambda_v + X_{ihv} + \nu_{ihv} \tag{2}$$

[14]The same specification can be used to examine caregiver-level outcomes — with the caveat that the i and h subscripts are redundant since only one primary caregiver is present in each household.

where L_{ihv} is an indicator for (household-level) random assignment to the Luo-language storybooks treatment and λ_v is a vector of village fixed effects. Again, we will include baseline values of the outcome variable whenever they are available. Since assignment to the Luo-language storybooks treatment occured at the caregiver level, standard errors will be clustered by household when estimating Equation 2.

We will construct confidence intervals using classical (asymptotic, frequentist) statistical methods. Our treatment assignment procedures involved re-randomization, but (as discussed above) we did not impose stringent constraints on the space of acceptable random assignments. As a result, randomization inference p-values should be comparable to or less conservative than those obtained from classical tests (Athey and Imbens 2017). Randomization inference p-values will be reported as a robustness check.

4.2 Defining the Endline Sample(s)

Our sample includes 2,013 caregivers and 2,527 children aged 36 to 83 months old at baseline. These individuals were surveyed pre-treatment, so a broad set of set of (baseline) covariates is available for both the **caregiver** sample and the **baseline child** sample. Through our baseline survey, we also identified 2,043 **older siblings** aged 84 to 143 months (at baseline) who were living in study households (at baseline) and 442 **younger siblings** who were 18 to 35 months old (at baseline). Though older and younger siblings were not assessed at baseline, they may be impacted by treatment — for example, older siblings may particularly benefit if they are asked to read with their younger brothers and sisters. Our endline data collection allows us to estimate impacts of treatment on all children aged 18 to 143 months at baseline — but, because of the wide age range and differential availability of baseline data, it will not always make sense to pool the samples in our analysis. Thus, our endline analysis involves four distinct sub-samples: the caregiver (or household-level) sample (N = 2,013), the baseline child sample (N = 2,527), the older sibling sample (N = 2,043), and the younger sibling sample (N = 442).

4.3 Hypotheses

4.3.1 Primary Hypotheses

Our primary research question is whether the EMERGE intervention improved children's vocabulary and early literacy skills. Though we expect vocabulary and literacy to be positively correlated, we consider these two classes of outcomes separately, in turn. We note that while we generally plan to estimate effects on age-normalized outcomes, there are at least two ways we can provide policy-relevant and easily-comparable scales for those standardized effect sizes. One is to compare an effect to the association between that age-normalized outcome and socioeconomic measures: how much it changes with mother's education, asset indices, *et cetera*. Another is to go back to raw (not age-normalized) scores, which we can benchmark against grade progression because we assess nearly all children in this study with a common set of instruments.[15]

Hypothesis 1. *The EMERGE treatment improved vocabulary outcomes among children in the baseline sample.*

We measure expressive and receptive vocabulary through direct child assessment. Receptive vocabulary refers to the ability to understand spoken words. Expressive vocabulary is the ability to produce appropriate words when required — for example, to name objects presented as images. Children begin developing receptive vocabulary before they begin to express themselves through speech (Fernald et al. 2017). We measure expressive vocabulary through a 57-item assessment developed and validated for the EMERGE study (Knauer et al. 2019b). It includes seven vocabulary words that were embedded in the intervention storybooks ("storybook expressive vocabulary") as well as 50 other locally-appropriate stimuli ("non-storybook expressive vocabulary"). Sample expressive vocabulary stimuli are shown in Appendix Figure A2. To assess receptive vocabulary in English and Luo, we developed and validated new assessments by adapting items from the British Picture Vocabulary Scale (a variant of the Peabody Picture Vocabulary Test suited to British or Commonwealth English), and by creating new, similarly-structured items appropriate to the local context (Dunn and Dunn 1997; Dunn, Dunn, and Styles 2009; Knauer et al. 2019b).

[15] All baselined children are presented with the same set of assessments; some older and younger siblings face only an age-appropriate subset.

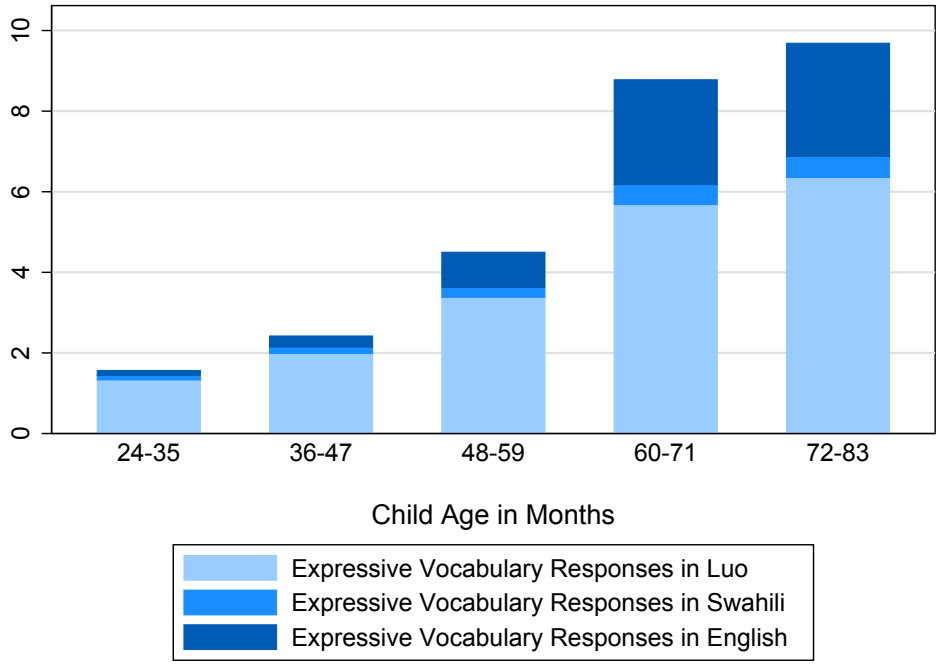
Our Luo and English receptive vocabulary assessments include 62 and 52 items, respectively. Sample receptive vocabulary stimuli are shown in Appendix Figure A3.

Dialogic reading interventions have been shown to have larger impacts on expressive vocabulary than on receptive vocabulary (Mol et al., 2008). Relative to most of the high-income-country contexts where dialogic reading has been evaluated, our setting is unusual because children are developing vocabulary skills in both their mother tongue and English. Figure 3, which is excerpted from Knauer et al. (2019b), illustrates the evolution of word choice as children in our study area age (using responses to an abbreviated expressive vocabulary assessment in our pilot study). The number of correct responses clearly increases with age as children learn more words. Young children use very few English or Swahili words; they express themselves almost exclusively in their mother tongue. However, older children — most of whom are enrolled in pre-primary or primary school — express themselves in a mix of Luo and English (though they still use very few Swahili words). Thus, expressive vocabulary provides a holistic measure of word knowledge because children can respond in any language — whereas receptive vocabulary assessments typically only measure knowledge of a single stimulus language at a time (Knauer et al. 2019b).

In multilingual settings, early literacy interventions can impact the trajectory of language development by nudging children to build skills in a particular language. Though existing evidence from monolingual contexts suggests that dialogic reading has larger impacts on expressive vocabulary than receptive vocabulary, our pilot study suggested that the EMERGE intervention may have had large impacts on receptive vocabulary in English. In Table 4, we report the estimated treatment effects of providing dialogic reading training and storybooks (including two in English) on expressive and receptive vocabulary (measured six-weeks after treatment in our pilot study). Treatment led to a 0.24 SD increase in knowledge of the vocabulary words embedded in the storybooks (p-value 0.036), and a 0.11 SD increase in an aggregate vocabulary index (p-value 0.096). Though confidence intervals are extremely wide in our ($N = 505$) pilot study sample, the point estimate suggests that the treatment effect on English receptive vocabulary may be quite large (0.15 SD, p-value 0.276, as shown in Table 4), and that impacts on non-storybook expressive vocabulary and receptive vocabulary in Luo may be considerable smaller.

In the present study, we will test the hypothesis that the EMERGE treatment improved chil-

Figure 3: The Use of Luo and English by Age — Results from Knauer et al. (2019b)



Notes: the figure is excerpted from Knauer et al. (2019b). It summarizes 505 children's responses to a 23-item expressive vocabulary assessment broken down by language. In the assessment, a child is shown an image — for example, a frog — and asked "What is this?" A child can then respond in the language of their choice; their response is marked as correct as long as it is an acceptable word for "frog" in English, Luo, or Swahili.

dren's vocabulary by estimating Equation 1. We consider four vocabulary outcomes (storybook expressive, non-storybook expressive, Luo receptive, and English receptive vocabulary), each expressed as an age-normalized z-score; we also construct an aggregate vocabulary index following the procedures outlined by Kling, Liebman, and Katz (2007). We consider the aggregate index in isolation, but treat the set of four component outcome variables as a family by adjusting p-values following the Benjamini-Hochberg procedure outlined by Anderson (2008).

When estimating the impacts of the EMERGE treatment on vocabulary, we restrict attention to the baseline child sample. There are two main reasons for doing this. First, baseline data on each of our four vocabulary outcomes is available for those children. As discussed further below, baseline and endline vocabulary measures were highly correlated in our pilot: the correlation between the baseline and endline vocabulary index was 0.796. Hence, including baseline values in our analysis should increase statistical power substantially — more so than increasing

the number of observations per cluster. Second, both our expressive vocabulary assessments and our Luo-language receptive assessment show ceiling effects among older children. For expressive vocabulary, it is well-known that assessments based on picture naming are not suitable for older children (Fernald et al. 2017). Though receptive vocabulary instruments (e.g. the Peabody Picture Vocabulary Test) are intended for use with both children and adults, in practice it proved impossible to identify difficult Luo words that displayed attractive psychometric properties (for example, discrimination, as measured through item response theory).[16] Ceiling effects limit one's ability to detect the impacts of treatment, so we restrict attention to the sub-sample where we have the best chance of detecting treatment effects on vocabulary. In addition to controlling for baseline values of the outcome variable, all specifications will include controls for child gender, baseline height-for-age z-score, and child age in months.

Hypothesis 2. *The EMERGE treatment improved children's early literacy skills.*

We measure early literacy skills using the English and Luo versions of the Early Grade Reading Assessment (EGRA). The EGRA measures four components of early literacy: knowledge of letter sounds, decoding, oral reading fluency, and reading comprehension (Dubeck and Gove 2015, RTI International 2015). We use existing (Kenya-specific) English and Luo versions of the EGRA with minimal adaptation, but we extend the test by including a letter recognition task adapted from the Malawi Developmental Assessment Test (Gladstone et al. 2010) as well as advanced reading passages adapted from Kenyan fifth-grade texts. Three of the four EGRA tasks (letter sounds, decoding, and fluency) are timed, yielding scores for both the number of correct responses and time to completion. We only assess advanced reading comprehension skills in children who complete the standard EGRA fluency and comprehension passages with minimal difficulty (no more than one incorrect response); reading comprehension scores are constructed by calculating the total number of correct responses across the easy and difficult passages. The early literacy index consequently contains 15 sub-components: letter recognition, knowledge of letter sounds (correct responses), time to complete letter sounds, non-word decoding (number correct), time to

[16] In piloting, we asked one woman about Luo words that her 13-year-old granddaughter did not know, but that Luo-speaking adults would know. The grandmother reported that her granddaughter "already knew all the Luo words." She went on to explain that there were many English words that her granddaughter hadn't learned yet, but that a smart 13-year-old already knew all the Luo words in common use.

complete non-word decoding, oral reading fluency (words read correctly from passage), time to complete oral reading fluency, and reading comprehension — all but the first of which is measured separately in English and Luo. We convert each component of the early literacy index into an age-normalized z-score using non-parametric procedures (Fan 1993). We will estimate the mean effect of the EMERGE treatment on early literacy by estimating Equation 1. In addition, we will estimate the impact of treatment on each outcome within the family, adjusting for the total number of tests following Anderson (2008).

We did not include a full early literacy module in our baseline survey. At baseline, we conducted direct assessments of children aged three to six years, most of whom had not yet begun learning to read. Our baseline survey included two brief measures of early literacy skills: the MDAT letter recognition task described above and a brief "familiar word reading" task wherein children were asked to read four simple words in English (bus, dog, cart, and sofa) and Luo (book, grass, mother, and cat — all of which are three or four letters in Luo). We did not measure early literacy in our short-term study either – again, because most children in our sample were too young for the EGRA. As a result, we have no way to predict the extent to which controlling for our brief pre-literacy skills assessment and other measures of early childhood development (for example, height-for-age z-score and expressive and receptive vocabulary at baseline) will explain the observed variance in literacy skills at endline. On the other hand, we observed only minimal ceiling effects when piloting our expanded EGRA with primary school children prior to endline — suggesting that the assessment is suitable for both our baseline child and older sibling samples.

This suggests two approaches to estimating the impacts of the EMERGE treatment on literacy skills. One option is to restrict attention to the baseline child sample ($N = 2{,}527$), including controls for baseline familiarity with letters, familiar word reading, height-for-age z-score, expressive and receptive vocabulary, age, and gender. The alternative is to pool the baseline child and older sibling samples ($N = 2{,}527 + 2{,}043 = 4{,}570$) even though many baseline covariates (pre-literacy skills, *et cetera*) are not available for the older sibling sample. Thus, if we pool the two samples, we would control for every child's age and gender, but only include other controls when available (including an indicator variable for whether the baseline child development controls are missing). The older sibling sample will have a higher residual variance, but may increase overall

sample size sufficiently to improve precision. Given the available data, there is no way to know in advance which estimate will have the smaller standard error. Instead of committing to one of the two specifications in advance, we commit to an explicit decision rule. Once endline data are available, we will calculate the variance of the endline outcome and the intra-class correlation for both proposed samples and sets of control variables; we will then test Hypothesis 2 by estimating the specification with the smaller minimum detectable effect (calculated following the procedures in Section 4.5). In either case, we estimate impacts by estimating Equation 1.[17]

4.4 Secondary Hypotheses

4.4.1 Luo vs. English Storybooks

In a first set of secondary hypotheses, we test for differential impacts of Luo-language and English-language storybooks on vocabulary and literacy outcomes. First, we compare each sub-treatment to the control group (Hypotheses 3 to 6), and then we compare the Luo-storybooks and English-storybooks treatments to each other (Hypotheses 7 and 8).

Hypothesis 3. *Luo books improved vocabulary among children in the baseline sample.*

Hypothesis 4. *Luo books improved children's early literacy skills.*

Hypothesis 5. *English books improved vocabulary among children in the baseline sample.*

Hypothesis 6. *English books improved children's early literacy skills.*

We will test Hypotheses 3, 4, 5, and 6 by estimating Equation 1. We follow the procedures outlined in Section 4.3.1 with respect to the selection of outcome variables, samples, and controls. However, when testing Hypotheses 3 and 4 (resp. Hypotheses 5 and 6), we restrict the sample to include only households assigned the control group and those assigned to the Luo-language storybooks (resp. English-language storybooks) treatment. Finally, we create two additional summary indices: early literacy in Luo and early literacy in English. Thus, we estimate the impact of each of two treatments (Luo storybooks and English storybooks) on each of four summary

[17] We also realize that the formulas in Section 4.5 may not exactly capture the variance structure in the presence of dummies for strata and so forth, so as a supplementary analysis, we will show the analysis both ways, for those interested in whether the power calculation led us to make the right decision.

outcomes (vocabulary, overall literacy, literacy in Luo, and literacy in English). We adjust p-values for multiple testing following Anderson (2008). There is clearly some double-counting — since outcomes used to construct the Luo and English literacy indices will also be used to construct the overall literacy index — creating a mechanical positive correlation among indices. We do this because it is critical to understand both the overall impact of each treatment on early literacy skills and their differential impacts on literacy in the storybook languages vs. other languages. In our supplementary analysis, we will also estimate the impact of each treatment (Luo and English storybooks) on each of the 19 component outcome variables used to construct our aggregate indices, adjusting p-values to reflect the fact that we conduct 34 hypothesis tests.

Hypothesis 7. *Luo and English books have different impacts on vocabulary among children in the baseline sample.*

Hypothesis 8. *Luo and English books have different impacts on children's early literacy skills.*

We will test Hypotheses 7 and 8 by estimating Equation 2. Again, we will consider the range of outcomes and samples described in Section 4.3.1. We will estimate the impact of the Luo storybooks treatment (as compared to the English storybooks treatment rather than the control group) on vocabulary, overall literacy, literacy in Luo, and literacy in English (as above), adjusting for multiple testing (as above). In our supplementary analysis, we will also estimate the impact of the Luo books treatment on each of the 19 component outcome variables used to construct our aggregate indices, adjusting p-values accordingly.

4.4.2 Mechanisms

A second set of secondary hypotheses tests a range of potential mechanisms that might explain any observed impact on child outcomes. Data collected from caregivers and children during our endline survey allows us to trace out an explicit causal chain: the EMERGE treatment increases the availability of age-appropriate reading materials in the home, causing parents and other family members to spend more time reading the EMERGE storybooks with their young children, which in turn leads to an increase in the overall level of stimulation young children experience. If we detect overall impacts in children's human capital, tracing out the steps in the causal chain

can help to explain the mechanisms underlying the observed results. If, on the other hand, we do not detect impacts on children's vocabulary and literacy skills, our analysis of intermediate outcomes will offer policy lessons about the design of early childhood interventions. Our midline survey documented clear impacts of treatment on the availability of books in the home and the frequency of shared reading. If these impacts do not persist, it indicates that relatively light-touch interventions such as ours are not sufficient to generate sustained changes in reading behaviors. In other words, the lack of children's books and information was not the binding constraint. However, if behavioral changes persist but do not translate into improvements in vocabulary and early literacy, it suggests that more intensive and costly early interventions may be required to generate durable improvements in the human capital of vulnerable children.

Hypothesis 9. *Treatment increases the number of children's storybooks in the home.*

Our midline analysis showed that the EMERGE treatment increased the likelihood that a household owned any children's storybooks by approximately 89 percent (p-value < 0.001), up from only 8 percent in the control group. However, if households do not value storybooks, they may be lost or destroyed over time. To test whether the EMERGE treatment increased the availability of children's reading materials 18 months post-treatment, we will estimate Equation 1 using both the indicator for the presence of any children's storybooks in the home and the number of children's storybooks in the home as outcome variables. In these specifications, analysis is at the caregiver/household-level and the only control is the baseline value of the outcome variable.[18][19]

Hypothesis 10. *Treatment increases the frequency of shared reading.*

We hypothesize that the EMERGE intervention could improve children's vocabulary and literacy because treatment increases the frequency with which older family members (either adults or older siblings) read with young children. Our endline survey provides several different measures of the frequency of shared reading:

[18] Additional control variables (e.g., mother's education or household assets) might improve statistical power. However, given the relative magnitudes of the coefficients and standard errors in our midline analysis of 364 households, we do not anticipate statistical power issues for analysis of this or (most) other intermediate outcomes.

[19] Also note that the field team is taking photos of any children's books found in homes at endline, so if desired, durability of the books could be assessed through a rubric for scoring the appearance of the books in the photos.

1. Caregivers were asked (directly) how often they read with each child over the past week. This provides ordinal measure of the frequency of caregiver-child book-sharing.

2. As part of an adapted version of the Family Care Indicators (FCI) questionnaire (Hamadani et al. 2010, Kariger et al. 2012), caregivers were asked whether *anyone* in the household read to or with each child in the three days prior to the survey. This generates seven indicators for family members who might have read with a child over the last three days (the child's mother, father, grandmother, grandfather, older sister, older brother, or anyone else).

3. As part of a time diary module developed and validated for the present study, caregivers were asked about all of their activities on the weekday prior to the survey. For each activity that a respondent engaged in, they report whether or not they were with their children and what (if any) activities they engaged in with their children — providing an indicator of whether the caregiver read with young children on the focus day for the time use module.

4. As part of the caregiver survey, we ask caregivers about the activities of older siblings — including whether or not older children read (or looked at books) with younger children.

5. As part of the child survey, older siblings are asked about their activities — including whether or not they read with any of their younger siblings.

Each of these measures has strengths and weaknesses. For example, direct questions about reading frequency may be subject to social desirability bias (since treatment-area respondents are aware that we are studying the storybook distribution program in which they participated). In contrast, older siblings' responses are unlikely to be influenced by social desirability bias (particularly social desirability bias that affects treatment and control groups differentially), but children's responses may be quite noisy. In our main analysis, we will construct an index of these five measures of reading frequency, following the procedures outlined by Kling, Liebman, and Katz (2007). In our supplementary analysis, we will estimate impacts on each of the outcomes, following Anderson (2008) to adjust for multiple hypothesis testing. We will also decompose the FCI measure of reading to the child into individual indicators: the indicator for whether the child's mother read with them in the last three days, the analogous indicator for the child's father, *et cetera*.

We test Hypothesis 10 by estimating Equation 1 for each child in the baseline child and younger sibling samples. Though outcome variables were collected through the caregiver and older sibling surveys, caregiver and older sibling responses characterize reading behaviors with each young child in the household. We will include the following child-level controls: child age, child gender, baseline height-for-age z-score (when available), baseline reading frequency. Only the second of our five measures of shared reading were measured at baseline, and only for children in the baseline sample. For those children, we will construct a baseline reading frequency index using the using only the first two measures of shared reading practices. For children in the younger sibling sample, we will construct a household-level average of the baseline reading frequency index to substitute for the child-specific baseline data.

Hypothesis 11. *Treatment increases children's familiarity with storybook content.*

Hypothesis 12. *Treatment increases primary caregivers' familiarity with storybook content.*

Hypothesis 13. *Treatment increases older siblings' familiarity with storybook content.*

Our endline child survey includes a series of comprehension questions that measure familiarity with storybook content by asking questions about illustrations taken from the stories. Figure A1 provides examples. Panel C of Figure A1, for example, is an illustration taken from the book *The Lovely Duck*; it shows two children looking through the grass at a duck building a nest. Though the duck and the nest are not shown, children (or adults) who have read the story will know the correct response. These questions measure the extent to which children and adults have used the storybooks enough to become familiar with their plots — a key step in the causal chain from book distribution to human capital impacts. In both our pilot study and our midline survey, we documented statistically significant and economically meaningful impacts of treatment on storybook comprehension among young children (in the baseline sample), demonstrating that children in treatment households use the EMERGE storybooks regularly in the first few months after treatment. At endline, we will measure impacts on both young children (in the baseline sample), older siblings, and caregivers by estimating Equation 1 for each sample.

Hypothesis 14. *Treatment increases the overall level of early childhood stimulation experienced by young children.*

Our adapted version of the Family Care Indicators (FCI) asks about ten different types of stimulating activities that adults and older siblings might engage in with young children: reading with them, telling them stories, singing to them, taking them places, playing with them, teaching them letters or numbers, counting with them, helping them learn new words (in either English or Luo), and helping them with homework. These can be used to create an overall index of stimulation experience by each young child, and they can be disaggregated to assess whether increases in shared reading are offset by declines in other types of stimulating activities. After each question in the FCI, we ask which household member engaged in the stimulating activity (mother, father, grandmother, grandfather, older sisters, older brothers, other adults, and/or other older children). This allows us to create indices of the amount of stimulation done by each (type of) household member — for example, the amount done by a child's mother as opposed to alloparents.

Our main analysis will estimate treatment effects on the overall amount of early childhood stimulation experienced by children in the baseline and younger siblings samples, controlling for child age and gender. We will estimate the overall impacts of the EMERGE treatment by estimating Equation 1. In our supplementary materials, we will also estimate the impact of treatment on the different types of stimulation experienced and on stimulation by each individual (type of) household member (mother, father, grandmother, grandfather, older sisters, older brothers, and other caregivers). All of these supplementary hypotheses will be treated as a family, with adjusted q-values calculated following Anderson (2008).

Hypothesis 15. *Treatment increases demand for additional children's storybooks.*

As a final measure of the impact of the EMERGE treatment on families' behaviors and attitudes related to book-sharing, we will estimate the impact of treatment on the demand for children's storybooks. After completing the caregiver survey, each respondent is offered a gift: respondents are given a choice between two large bars of soap or a (new) children's storybook. Our piloting (in non-study communities) suggests that households that have received storybooks in the past are approximately 30 percentage points more likely to choose additional books as their respondent gift. In our endline analysis, we will formally test this by estimating Equation 1 in the caregiver sample, controlling for baseline caregiver literacy, maternal education, and durable assets.

We will treat Hypotheses 9 through 15 as a family, again correcting for multiple hypothesis

tests following Anderson (2008). As discussed above, we will also estimate treatment effects on the likelihood that different family members have read with young children in the past three days (in supplementary analysis). Along with analysis of storybook comprehension among caregivers and older siblings, this will help us understand how households optimize their response to this literacy intervention.

4.4.3 Impacts on Older Siblings

Our baseline data demonstrates that older sisters do more early childhood stimulation than anyone else in the household, yet their role in mediating the impacts of early childhood ("parenting") interventions is typically ignored. Our endline data collection allows us to estimate the impacts of treatment on children aged 84 to 143 months (i.e. 7 to 11 years) at baseline. As discussed above, these children may benefit from the increased availability of early reading materials in the home, particularly if they engage in shared reading with younger siblings. On the other hand, emphasizing the importance of early childhood stimulation to parents may cause them to shift attention toward preschoolers — and away from school-aged children. Older children might also be asked to do more early childhood stimulation themselves, or to assist (more) with other household chores to free up parental time for shared reading and early childhood stimulation.

We consider the following outcomes:

1. Literacy in English, measured using the expanded EGRA (described above)
2. Literacy in Luo, also measured using the expanded EGRA (described above)
3. Receptive vocabulary in English, measured using the adapted British Picture Vocabulary Scale (described above)
4. School attendance the weekday prior to the survey
5. An indicator for whether the sibling reports reading with any younger siblings on the weekday prior to the survey
6. An indicator for whether the sibling reports reading on their own on the weekday prior to the survey

7. An indicator for whether the sibling did any homework on the weekday prior to the survey

8. An indicator for whether the sibling received any help with homework on the weekday prior to the survey

9. A chore index that tallies the number of different chores the older sibling did on the weekday prior to the survey

We treat these outcomes as a family, estimating impacts on each outcome and adjusting for multiple hypothesis testing following Anderson (2008). We assess the impacts of treatment on older siblings by estimating Equation 1 in the older sibling sample. We control for child age and gender, as well as the number of older and young male and female siblings in the household at baseline. Baseline values of the outcome variables are not available for the older sibling sample, so we cannot include them as controls. Since sisters play a much larger role in looking after young children then brothers (Lancy 2015), we estimate heterogeneous treatment effects on older girls versus older boys for all outcomes in this family (in addition to pooled impacts on older siblings).

4.4.4 Impacts on Primary Caregivers

In a final set of secondary hypotheses, we estimate the impacts of treatment on primary caregivers' literacy and time use. We estimate impacts of treatment on caregiver literacy in English and in Luo, and on time spent: (a) with young children; (b) stimulating young children; (c) on social leisure; (d) on non-social leisure and rest; (e) on home production, (f) on domestic work and household chores, and (g) on income-generating activities. All time use outcomes are measured through an open-interval time diary that was developed and validated for the EMERGE project. We estimate treatment effects via Equation 1 controlling for whether the primary caregiver is the children's mother, caregiver literacy (at baseline), and a household asset index.

4.5 Power Calculations

As a preable to (and overview of) the calculations below, there are a few central features of this study worth pointing out. In preparing these calculations, we are able to draw not only on the full baseline data from the present study, but also on data from our pilot study (Knauer et al. 2019),

which used measures analogous or identical to those we use in this study. These two sources of information provide the statistical parameters needed for our power calculations. This is a cluster-randomized trial; statistical power (equivalently, the minimum detectable effect) depends not only on sample size, but also on the number and size of clusters, the intra-class correlation of the outcomes, and the predictive power of baseline data. For most child outcomes, intra-class correlations are relatively small after conditioning on baseline values of the outcome variable, which are available for our primary child-level outcomes of interest. In many cases, we anticipate having power to detect 0.1-standard-deviation effects.[20] We discuss in more detail below.

Two other issues are typically considered when calculating statistical power: compliance (take-up) and attrition. As discussed in the previous sections, nearly every caregiver assigned to treatment received the intervention or some part of the intervention (specifically, the storybooks). Six weeks after treatment, storybooks were still present in respondents' homes and children in the treatment group had learned the stories while, unsurprisingly, those in the comparison group had not. Because of this, we do not consider compliance or take-up to be a concern here.

In terms of the magnitude of attrition, our team is following households with young children in a rural area, approximately 1.5 years after baseline; we have extensive contact information that allows us to track respondents through extended family, even if they move. We therefore expect attrition from household moves to be very low. We now provide two points of reference on attrition in this context. First, in a neighboring area of rural Kenya, teams from Innovations for Poverty Action (with which we are working on the present study) followed respondents ten years after initially enrolling them in a study, and were able to find and survey over 82 percent of them (Baird et al. 2016); given that our follow-up period is less than two years as opposed to ten, we expect to be able to locate a much higher fraction of baseline households. Second, for a sample of young women on the outskirts of Nairobi, a team hired and managed by some of the authors of the present study (again, in partnership with Innovations for Poverty Action) was able to track and survey over 92 percent of respondents after a year and a half (Brudevold-Newman et al. 2017). The EMERGE sample is substantially less mobile than that one, so we again expect the situation to allow us to do better. In light of these benchmarks, we think it is reasonable to

[20] A recent review found that interventions which, like ours, aim to train caregivers on responsive behaviors, typically had relatively large effects, somewhat allaying power concerns in general (Prado, et al. 2019).

anticipate that we will be able to follow up at least 90 percent of our respondents.

Moreover, individual-level attrition in a cluster-randomized study is not very problematic. In particular, the reduction in sample size causes the most severe adjustments to minimum detectable effects (MDE) when the intra-cluster correlation is zero; but this is also the condition under which we have the most power to begin with. In our setting, when intra-cluster correlation is zero, 10 percent attrition only increases MDEs by about 5.4 percent; when it is non-zero, MDEs only increase by about 1.4 percent. We explain the underlying formulas in Section 4.5.5, but for the remainder of this section, we focus on the full sample, as these adjustments are small.

4.5.1 Characterizing the Power Calculation Formula

The Minimum Detectable Effect (MDE) in a cluster-randomized trial is given by:

$$\text{MDE} = (t_{1-\kappa} + t_{\alpha/2})\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}\sqrt{1 + (n_{groupsize} - 1)\rho}. \tag{3}$$

The MDE expresses the statistical power of a test in standard deviations of the outcome variable. In what follows, we normalize $\sigma = 1$ and express the MDE in σ units.

Any power calculation of this sort involves the sum of two values of the T (or Z) distribution, which we use to obtain an approximation of $t_{1-\kappa} + t_{\alpha/2}$. The first component is the critical value at which a test rejects — this is a function of test size, and we usually consider a two-sided test of size $\alpha = 0.05$. The second component is the value of the T (or Z) distribution at which the mass to the left of that point in the distribution equals the desired power (typically $\kappa = 0.8$). When $\alpha = 0.05$ and $\kappa = 0.8$, the result is approximately 2.8.[21]

We re-write the formula for the MDE as follows:

$$\text{MDE} \approx 2.8\left(\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{1}{N}}\sqrt{1 + (n_{groupsize} - 1)\rho}\right). \tag{4}$$

[21] One could use the t-distributed variant with 1000 degrees of freedom (a figure just under the number of caregivers in the treatment arms): $t_{1000,0.975} + t_{1000,0.8} =$ `invttail(1000,0.025)+invttail(1000,0.2) = 2.80431...` . Alternatively, for analysis where treatment varies at the community level rather than within-community, the t-distributed variant with 71 degrees of freedom (the number of clusters less two degrees of freedom) may be more appropriate: $t_{71,0.975} + t_{71,0.8} =$ `invttail(71,0.025)+invttail(71,0.2) = 2.84065...` . These differ only in the second decimal place, so any attrition at endline may do more to change the power calculations than switching among these variants of the power calculation formula.

N and P will vary across hypotheses, depending on the estimation sample being used and the test being conducted. Our study is cluster-randomized at the community level with 73 clusters, so $n_{groupsize} = N/73$. Since storybook language was randomized at the caregiver rather than the community level, analysis comparing the impacts of the English and Luo storybook treatments can be clustered at the household rather than the community level.

In any cluster-randomized study, power depends on the intra-cluster correlation, ρ, which varies across outcomes and samples. Power increases when we are able to control for a baseline value of an outcome variable — to the extent that the baseline value predicts the endline value, absorbing unexplained variance. Other baseline covariates – for example, child age, gender, and height-for-age z-score — may also predict endline outcomes and increase power. Specifically, for an outcome variable, Y, and a vector of baseline covariates, X, we can write the MDE as:

$$\text{MDE} \approx 2.8 \left(\sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{1}{N}} \sqrt{1 + (n_{groupsize} - 1)\tilde{\rho}} \right) \tilde{\sigma} \tag{5}$$

where $\tilde{\sigma} < 1$ is the standard deviation of the residual of Y after regressing it on X and $\tilde{\rho}$ is the intra-class correlation of the residuals. We would typically expect $\tilde{\rho} < \rho$, but this might not always be the case.[22]

We obtain estimates of ρ from our baseline data whenever possible. For variables that were not measured at baseline, we rely on proxy variables from our baseline survey, the short-term evaluation of the EMERGE intervention, or other studies. When considering outcomes where baseline data are available, we estimate $\tilde{\rho}$ and $\tilde{\sigma}$ using baseline and endline data from the short-term study whenever possible. Estimates of $\tilde{\rho}$ and $\tilde{\sigma}$ allow us to calculate an **adjusted MDE** accounting for covariates. We walk through the calculation of these quantities for our primary hypotheses, then summarize assumptions and MDEs for other outcomes in Table 5.

4.5.2 Primary Hypotheses

Hypothesis 1. Hypothesis 1 is that the EMERGE treatment improved vocabulary outcomes of children in the baseline sample. For this test, $N = 2{,}527$, and approximately half of the sample

[22] An obvious situation when we would expect $\tilde{\rho}$ to be greater than ρ is when an unobserved cluster-randomized treatment explains most of the change in the outcome variable over time.

was assigned to treatment ($P = 0.5$). We can thus re-write the MDE as:

$$\begin{aligned} \text{MDE}_1 &\approx \left((t_{1-\kappa} + t_{\alpha/2}) \cdot \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{1}{N}} \sqrt{1 + (n_{groupsize} - 1)\rho} \right) \\ &\approx \left(2.8 \cdot \sqrt{\frac{1}{(1/2)(1/2)}} \cdot \sqrt{\frac{1}{2527}} \sqrt{1 + \left(\frac{2527}{73} - 1 \right) \rho} \right) \end{aligned} \tag{6}$$

Our index of vocabulary skills averages four outcome variables: storybook-specific expressive vocabulary, (non-storybook) expressive vocabulary, receptive vocabulary in Luo, and Receptive vocabulary in English. We are also interested in estimating treatment effects on each of these outcomes independently. Baseline values of all four outcomes are available for all children in the baseline sample. This allows us to use the intra-class correlation (at baseline) to approximate ρ, and it means that we can use baseline values to increase that statistical power of our analysis.

First, consider storybook-specific expressive vocabulary, which had the highest baseline intra-class correlation ($\rho \approx 0.076$).[23] In the absence of baseline data, our research design yields an MDE of 0.213 SD — which we can confirm using the `sampsi` command in Stata:

```
sampsi 0 0.213, power(0.8) alpha(0.05) ratio(1) sd(1)
sampclus, numclus(73) rho(0.076)
```

Next, we use baseline and endline data from the short-term study to estimate the serial correlation for this specific outcome. As expected, regressing age-normalized storybook expressive vocabulary on a constant generates residuals with a standard deviation of approximately 1. Adding baseline values of the outcome variable to the regression reduces the standard deviation of the residuals to approximately 0.77, indicating a $\tilde{\sigma}$ of approximately 0.77. This alone would decrease the MDE to approximately 0.163 — as we can confirm in Stata:

```
sampsi 0 0.163, power(0.8) alpha(0.05) ratio(1) sd(0.77)
sampclus, numclus(73) rho(0.076)
```

However, including baseline values of the outcome variable also reduces the intra-class correlation to 0.026 (in the short-term pilot evaluation) — suggesting an adjusted MDE of 0.119.

[23] Throughout, we calculate intra-class correlations using data from children aged four to six at baseline — to approximate the endline age distribution as closely as possible.

In the first five rows of Table 5, we replicate these calculations for our other three vocabulary measures and our aggregate vocabulary index. Without baseline covariates, MDEs vary with the intra-class correlation (because the size of the sample, the proportion treated, and other aspects of the research design are held constant). The unadjusted MDE is lowest (at 0.140) for Luo receptive vocabulary because it has the lowest intra-class correlation in our baseline data. However, data from the pilot study demonstrates that the intra-class correlation is always substantially lower after conditioning on baseline values of the outcome variable, and our covariate-adjusted MDEs range from 0.084 to 0.119. We observe the lowest covariate-adjusted MDE for our aggregate vocabulary index. Importantly, the aggregation process also reduces the overall level of noise in this outcome, so the ability to detect a 0.084 SD impact on vocabulary represents a high degree of power to detect small impacts on child vocabulary.

Hypothesis 2. Our second primary hypothesis is that the EMERGE treatment improved children's literacy skills. As discussed above, we do not have baseline values of the outcome variable, though we have measures of performance on a (relatively brief) familiar word reading task for children in the baseline sample. The question is whether we increase power by including the older sibling sample — for whom baseline measures of height-for-age and familiar word reading are not available. If the baseline covariates are sufficiently predictive, the increase in precision resulting form the larger sample could be offset by the increase in unexplained variance.

In Table 5, we report the unadjusted MDE for both samples. As expected, the unadjusted MDE is lower (at 0.165) for the larger sample. However, the relatively high intra-class correlation in baseline familiar word reading skills ($\rho = 0.046$) means that including the older siblings has only a modest impact on the MDE (since the increase in sample size does not increase the number of clusters). Thus, if baseline covariates do a good job or predicting EGRA scores at endline, the restricted baseline sample could easily yield a smaller MDE.

4.5.3 Secondary Hypotheses

Hypotheses 3, 4, 5, and 6 test the impacts of the Luo-language and English-language storybooks treatments relative to the control group. Since storybook language was randomly assigned within the treatment group, such analysis generates unbiased estimates of treatment effects. However,

the treatment clusters are only half as large as the control clusters. We therefor begin our discussion of our secondary hypotheses by deriving the slightly-different MDE formula when cluster size differs between study arms. We then summarize the MDEs for our secondary hypotheses.

Uneven Clusters. Here we discuss power when treatment and comparison groups have differently-sized clusters, as a simple extension to the usual formula. First, recall that the minimum detectable effect is of the form

$$\text{MDE} = (t_{1-\kappa} + t_{\alpha/2})\widehat{SE}, \tag{7}$$

as shown previously in Equation 3. The question is how to calculate $\widehat{SE}$. This is the standard error from a difference between two means. Recall that there may be intra-cluster correlation. That is, underlying each residual error term $\tilde{y}_i$, we have $\tilde{y}_i = \epsilon_i + \eta_g$, where: ϵ_i are iid with variance σ^2_ϵ; η_g are iid across groups but identical within groups, with variance σ^2_η; thus the variance of $\tilde{y}_i$ equals $\sigma^2 = \sigma^2_\epsilon + \sigma^2_\eta$. The intra-cluster correlation, ρ, equals $\sigma^2_\eta/(\sigma^2_\eta + \sigma^2_\epsilon)$ (equivalently, σ^2_η/σ^2). Then it is straightforward to calculate the variance of a group mean for a subsample of size s_1 with groups of size g_1 and intra-cluster correlation ρ:

$$Var\left(\frac{1}{s_1}\sum_{s_1}\tilde{y}_i\right) = \frac{1}{s_1^2}Var\left(\sum_{s_1}(\epsilon_i + \eta_g)\right) = \frac{1}{s_1^2}\left(s_1\sigma^2_\epsilon + g_1^2 \cdot \frac{s_1}{g_1}\sigma^2_\eta\right) = \frac{1}{s_1}\left(\sigma^2_\epsilon + g_1\sigma^2_\eta\right).$$

The right term follows from the facts that (a) for groups of n_g observations, the η_g terms are identical, so the variance of the sum of η_g is the square of the number of summed terms times the underlying variance, and (b) that for a subsample of size s_1, the number of such groups is s_1/g_1. Simplifying the above expression (and using the definition of ρ), we find:

$$\frac{1}{s_1}\left(\sigma^2_\epsilon + g_1\sigma^2_\eta\right) = \frac{1}{s_1}\left((\sigma^2_\epsilon + \sigma^2_\eta) + (g_1 - 1)\sigma^2_\eta\right) = \frac{1}{s_1}\sigma^2(1 + (g_1 - 1)\rho). \tag{8}$$

The standard error we want will be the square root of the variance of a difference between two means. Those means involve mutually exclusive groups of potentially different size. Thus:

$$\widehat{SE} = \sigma \cdot \sqrt{\frac{1}{s_1}(1 + (g_1 - 1)\rho) + \frac{1}{s_2}(1 + (g_2 - 1)\rho)}. \tag{9}$$

When $s_1 = Pn$, $s_2 = (1-P)n$, and $g_1 = g_2 = n_{groupsize}$, it is readily seen that this simplifies to the well-known formula shown earlier (in Equation 3 and elsewhere). But the formula in Equation 9 governs the more general case, a relevant instance of which we now describe.

Luo vs. English Storybooks When testing Hypotheses 3, 4, 5, and 6, we compare the mean from half the treatment group to that from the entire comparison group. Following Equation 9 above, we have $s_1 \approx 635$, $g_1 \approx 635/36 \approx 17.7$, $s_2 = 1260$, and $g_2 = 1260/37 \approx 34.054$.[24]

We discuss power in the case of our aggregate vocabulary index ($\rho \approx 0.060$). In the absence of baseline data, our research design yields an MDE of approximately 0.208. Adding baseline values of the outcome variable to the regression reduces the standard deviation of the residuals to approximately 0.77; this alone would decrease the MDE to approximately 0.126. However, including baseline values of the outcome variable also reduces the intra-class correlation to approximately 0.015 (in the pilot evaluation sample) — suggesting an adjusted MDE of 0.095.

As discussed above, we consider two approaches to estimating treatment effects on literacy — either restricting attention to the baseline sample or pooling the baseline and older sibling samples. As expected, the larger sample yields the smaller MDE in the absence of baseline data. Without baseline data, we get an MDE of 0.159 by pooling the two samples; we may be able to lower the MDE by including our baseline measures of child development and literacy.

Hypotheses 7 and 8 compare the impacts of the Luo-language and English-language storybooks.[25] Since storybook language was randomized at the household level, we cluster accordingly and include community fixed effects in our analysis. Baseline and pilot data indicate that the within-household intra-class correlation is typically quite high for vocabulary and literacy outcomes, but this has limited impacts on power since the number of observations per cluster is very small (e.g. 1.25 children per household in the baseline sample). Thus, unadjusted MDEs are in the 0.145–0.165 range, and the covariate-adjusted MDE indicates that we are powered to detect vocabulary impacts as small as 0.096 standard deviations.

[24] There are 635 children assigned to receive Luo language books, and 632 children assigned to receive English language books.

[25] Since Hypotheses 5 and 6 replicate Hypotheses 3 and 4 for the English (rather than Luo) books treatment, we do not discuss the power calculations in detail — they are identical to those described above except for tiny differences in sample size.

Mechanisms. Hypotheses 9 through 15 consider a range of mechanisms: impacts of treatment on the presence of books in the home, on the frequency of shared-reading, on familiarity with storybook content, and on the demand for storybooks. Most of these outcomes were not measured at baseline. However, as Table 3 demonstrates, we anticipate large treatment effects since these are, in effect, process outcomes. For example, our midline survey found a treatment effect of 8.7 SDs on the number of storybooks in the home — relative to an estimated MDE of 0.150. At midline, we found a treatment effect of 6.5 SDs on our measures of familiarity with storybook content — relative to an unadjusted MDE of 0.269. Thus, we are well-powered to detect likely impacts on intermediate outcomes, even without variance-reducing baseline data.

Impacts on Caregivers and Siblings. We consider a range of secondary hypotheses testing impacts of treatment on older household members. For most of these outcomes (caregiver literacy being a noteable exception), we do not have baseline values of the outcome variables of interest. Consequently, we do not walk through these secondary hypotheses in great detail. In Table 5, we summarize the range of MDEs that we can expect across a plausible space of intra-class correlations given our sample sizes and research design. We are generally well-powered to detect moderate impacts in the 0.1–0.25 SD range.

4.5.4 Heterogeneous Effects

In our pilot study, we found that treatment effects were generally larger for the children of illiterate caregivers (Knauer et al. 2019a). To see whether this pattern replicates, we will test for treatment effect heterogeneity by (baseline) caregiver literacy, but only for our primary hypotheses. As discussed above, we also report gender-disaggregated analysis of impacts on older siblings (in addition to pooled analysis). Finally, we will present exploratory analysis disaggregating treatment effects on vocabulary and literacy by child age. In all cases, we will adjust p-values for multiple hypothesis testing following Anderson (2008).

4.5.5 Adjustment for Attrition

When adjusting for attrition, straightforward reasoning holds that if attrition is not differential, minimum detectable effects will be adjusted by the inverse of the square root of the follow-up

rate.[26] That is, if 90 percent of respondents are found, the sample size shrinks by a factor of 0.9, causing the minimum detectable effect (MDE) to rise by a factor of $1/\sqrt{0.9} \approx 1.054$, about a five percent increase in MDE.

This is true when the intra-cluster correlation of the outcome (ρ) is zero, which is of course also the condition under which the MDE is smallest (the power is highest) to begin with. However, when $\rho \neq 0$, the adjustment for attrition is slightly different, because while the sample size does change, the "design effect" or "moulton factor" changes in a countervailing way. This, in turn, is because we anticipate attrition not to occur at the cluster level, but at the individual level: entire clusters are unlikely to be lost to follow-up, but hard-to-find individuals within those clusters may attrit. Thus, the number of clusters is constant, but the number of observations per cluster falls. Let the number of clusters be $G = N/n_{groupsize}$. Then, because we can express $n_{groupsize}$ as N/G, we can build on Equation 5:

$$\frac{MDE_{attrition}}{MDE_{full}} \propto \sqrt{\frac{N_{full}}{N_{attrition}}} \sqrt{\frac{(1 + ((N_{attrition}/G) - 1)\rho}{(1 + ((N_{full}/G) - 1)\rho}}. \tag{10}$$

In a typical power scenario for the present study, we have $N_{full} = 2527$, and $G = 73$. Suppose $N_{attrition}/N_{full} = 0.9$ (ninety percent follow-up, ten percent attrition). In some cases, we have $\rho = 0$, in which case the formula for adjusting the MDE in Equation 10 reduces to only the left term (since the radical on the right equals one), yielding $\sqrt{N_{full}/N_{attrition}} = \sqrt{1/0.9} \approx 1.054$, as discussed above. However, we also expect cases in which intra-cluster correlation is nonzero, such as $\rho = 0.076$. In this case, the radical on the right of Equation 10 substantially counteracts the radical on the left, yielding $1.054 \cdot 0.962 \approx 1.014$. Intuitively, when $\rho \neq 0$, as group size grows very large, additional observations within the group no longer affect the standard error. Thus, even ten percent individual-level attrition may only change our MDE by one percent.

[26] Differential attrition raises the question of approaches to bounding. Because we anticipate very low attrition overall, we do not discuss differential attrition here.

5 Administrative Information

5.1 Funding

This research project is supported by Echidna Giving, and by the World Bank via: the Strategic Impact Evaluation Fund; the Early Learning Partnership; and the Research Support Budget.

5.2 Institutional Review Board

Necessary approvals are in place. This study has been reviewed and approved by three entities: by the University of California at Berkeley's Committee for the Protection of Human Subjects (under Protocol ID 2014-09-6699); in Kenya, by the Maseno University Ethics Review Committee (under Proposal Reference Number MSU/DRPC/MUERC/00118/14); and finally, by the Innovations for Poverty Action Institutional Review Board (under Protocol Number 2620).

5.3 Declaration of Interest

The authors state that they have no competing conflicts.

5.4 Acknowledgements

We are grateful to Andrew Brudevold-Newman, William Blackmon, Rohit Chhabra, Mathilda Chweya, Emily Cook-Lundgren, Julian Duggan, Sheyda Esnaashari, Gerald Jona Ipapa Etarukot, Patricia Gitonga, Jessica Jomo, Saahil Karpe, Laura Kincaide, and Elyse Thulin for research assistance. We are particularly grateful to Patricia Kariger and Frances Aboud for their collaboration with us on the earlier papers that make the present study possible. Many others, including Sarah Baird, Emanuela Galasso, Guthrie Gray-Lobe, Joan Hamory Hicks, Alaka Holla, Anthony Keats, Michael Kremer, Isaac Mbiti, and Edward Miguel, provided helpful comments and access to previously-developed survey instruments. This document is a revision to a registered report initially submitted to the Journal of Development Economics on August 31, 2019.

References

Alderman, Harold, Jere R Behrman, Paul Glewwe, Lia Fernald, and Susan Walker, "Chapter 7 - Evidence of impact of interventions on growth and development during early

and middle childhood," in DAP Bundy, ND Silva, S Horton, DT Jamison, and GC Patton, eds., *Disease Control Priorities, (Volume 8): Child and Adolescent Health and Development, 3rd edition*, World Bank Publications, 2017, p. 1790.

Almond, Douglas and Janet Currie, "Chapter 15 - Human capital development before age five," in David Card and Orley Ashenfelter, eds., *Handbook of Labor Economics*, Vol. 4, Elsevier, 2011, pp. 1315 – 1486.

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

Andrew, Alison, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina, "Impacts 2 years after a scalable early childhood development intervention to increase psychosocial stimulation in the home: A follow-up of a cluster randomised controlled trial in Colombia," *PLoS Medicine*, 2018, *15* (4), e1002556.

Athey, Susan and Guido W. Imbens, "The econometrics of randomized experiments," in Esther Duflo and Abhijit Banerjee, eds., *Handbook of Economic Field Experiments*, Vol. 1, Elsevier, 2017, pp. 73–140.

Attanasio, Orazio P., Camila Fernández, Emla O. A. Fitzsimons, Sally M. Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina, "Using the Infrastructure of a Conditional Cash Transfer Program to Deliver a Scalable Integrated Early Child Development Program in Colombia: Cluster Randomized Controlled Trial," *BMJ*, 2014, *349.*

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel, "Worms at work: Long-run impacts of a child health investment," *The Quarterly Journal of Economics*, 2016, *131* (4), 1637–1680.

Ball, J., "Enhancing learning of children from diverse language backgrounds: Mother tongue-based bilingual or multilingual education in the early years," Technical Report 2010.

Behrman, J. R., J. Hoddinott, J. A. Maluccio, E. Soler-Hampejsek, E. L. Behrman, R. Martorell, Manuel Ramirez-Zea, and A. D. Stein, "What Determines Adult Cognitive Skills? Influences of Pre-School, School, and Post-School Experiences in Guatemala," *Latin American Economic Review*, 2014, *23* (4).

Black, Maureen M, Susan P Walker, Lia CH Fernald, Christopher T Andersen, Ann M DiGirolamo, Chunling Lu, Dana C McCoy, Günther Fink, Yusra R Shawar, Jeremy Shiffman et al., "Early Childhood Development Coming of Age: Science through the Life Course," *Lancet*, 2017, *389* (10064), 77–90.

Brudevold-Newman, Andrew Peter, Maddalena Honorati, Pamela Jakiela, and Owen W. Ozier, "A Firm of One's Own: Experimental Evidence on Credit Constraints and Occupational Choice," World Bank Policy Research Working Paper 7977 2017.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232.

Carneiro, Pedro and James Heckman, "Human Capital Policy," Working Paper 9495, National Bureau of Economic Research February 2003.

Chicoine, Luke, "Schooling with learning: The effect of free primary education and mother tongue instruction reforms in Ethiopia," *Economics of Education Review*, 2019, *69*, 94–107.

Commission on Revenue Allocation, "Kenya County Fact Sheets," Produced by the Commission on Revenue Allocation, Australian AID, Kenya National Bureau of Statistics, and the World Bank 2011.

Das, Jishnu, Stefan Dercon, James Habyarimana, Pramila Krishnan, Karthik Muralidharan, and Venkatesh Sundararaman, "School Inputs, Household Substitution, and Test Scores," *American Economic Journal: Applied Economics*, 2013, *5* (2), 29–57.

Dillon, Moira R., Harini Kannan, Joshua T. Dean, Elizabeth S. Spelke, and Esther Duflo, "Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics," *Science*, 2017, *357* (6346), 47–55.

Dubeck, Margaret M. and Amber Gove, "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations," *International Journal of Educational Development*, 2015, *40*, 315–322.

Dunn, L. M. and D. M. Dunn, "PPVT-III: Peabody Picture Vocabulary Test," 1997.

___ , ___ , and B. Styles, "British Picture Vocabulary Scale (3rd ed.)," 2009.

Duursma, Elisabeth, Marilyn Augustyn, and Barry Zuckerman, "Reading Aloud to Children: The Evidence," *Archives of Disease in Childhood*, 2008, *93* (7), 554–557.

Elmonayer, R. A., "Promoting phonological awareness skills of Egyptian kindergarteners through dialogic reading," *Early Child Development and Care*, 2013, *183* (9), 1229–1241.

Fernald, Lia C. H., Elizabeth Prado, Patricia Kariger, and Abbie Raikes, "A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries," Technical Report, International Bank for Reconstruction and Development / The World Bank, Washington, DC 2017.

Filmer, Deon and Lant H. Pritchett, "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollments in States of India," *Demography*, 2001, *38* (1), 115–132.

Galasso, Emanuela, Ann Weber, and Lia CH Fernald, "Dynamics of child development: Analysis of a longitudinal cohort in a very low income country," *The World Bank Economic Review*, 2019, *33* (1), 140–159.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor, "Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica," *Science*, 2014, *344* (6187), 998–1001.

Gladstone, M., G. A. Lancaster, E. Umar, M. Nyirenda, E. Kayira, N. R. van den Broek, and R. L. Smyth, "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* (5), http://doi.org/10.1371/journal.pmed.1000273.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin, "Many children left behind? Textbooks and test scores in Kenya," *American Economic Journal: Applied Economics*, 2009, *1* (1), 112–35.

Grantham-McGregor, Sally, Yin Bun Cheung, Santiago Cueto, Paul Glewwe, Linda Richter, Barbara Strupp, International Child Development Steering Group et al., "Developmental potential in the first 5 years for children in developing countries," *The Lancet*, 2007, *369* (9555), 60–70.

Hamadani, J. D., F. Tofail, A. Hilalay, S. N. Huda, P. Engle, and S. M. Grantham-McGregor, "Use of Family Care Indicators and Their Relationship with Child Development in Bangladesh," *Journal of Health, Population, and Nutrition*, 2010, *28* (1), 23–33.

Hanushek, Eric A and Ludger Woessmann, "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation," *Journal of Economic Growth*, 2012, *17*, 267–321.

High, Pamela C, Linda LaGasse, Samuel Becker, Ingrid Ahlgren, and Adrian Gardner, "Literacy promotion in primary care pediatrics: can we make a difference?," *Pediatrics*, 2000, *105* (Supplement 3), 927–934.

Jones, J.M., "The Effect of Language Attitudes on Kenyan Stakeholder Involvement in Mother Tongue Policy Implementation," *Journal of Multilingual and Multicultural Development*, 2012, *33* (3), 237–250.

Jukes, M. C. H., P. Gabrieli, N. L. Mgonda, F. S. Msolezi, G. Jeremiah, J. J. Tibenda, and K. L. Bub, "Respect Is an Investment: Community Perceptions of Social and Emotional Competencies in Early Childhood from Mtwara, Tanzania," *Global Education Review*, 2018, *5* (2), 160–188.

Kariger, P., E. A. Frongillo, P. Engle, P. M. Britto, S. M. Sywulka, and P. Menon, "Indicators of Family Care for Development for Use in Multicountry Surveys," *Journal of Health, Population, and Nutrition*, 2012, *30* (4), 472–486.

Kenya National Bureau of Statistics, "Nyanza Province Multiple Indicator Cluster Survey 2011, Final Report," Nairobi, Kenya: Kenya National Bureau of Statistics 2013.

Kerwin, Jason T. and Rebecca L. Thornton, "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures," mimeo 2019.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, "Experimental Analysis of Neighborhood Effects," *Econometrica*, 2007, *75* (1), 83–119.

Knauer, Heather A., Pamela Jakiela, Owen Ozier, Frances Aboud, and Lia C. H. Fernald, "Enhancing young children's language acquisition through parent-child book-sharing: a randomized trial in rural Kenya," *Early Childhood Research Quarterly*, 2019.

___ , **Patricia A. Kariger, Pamela Jakiela, Owen Ozier, and Lia C. H. Fernald**, "Multilingual Assessment of Early Child Development: Analysis from Repeated Observations of Children in Kenya," *Developmental Science*, 2019.

Lancy, David F., *The Anthropology of Childhood: Cherubs, Chattel, Changelings*, Cambridge, UK: Cambridge University Press, 2015.

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin, "Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools," mimeo 2020.

LeVine, Robert A, Suzanne Dixon, Sarah LeVine, Amy Richman, Constance H Keefer, P Herbert Leiderman, and T Berry Brazelton, *Child care and culture: Lessons from Africa*, Cambridge University Press, 1996.

Lewis, M.P., G.F. Simons, and C.D. Fenning, *Ethnologue: Languages of the World*, Dallas, TX: SIL International, 2016.

Lucas, Adrienne M, Patrick J McEwan, Moses Ngware, and Moses Oketch, "Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda," *Journal of Policy Analysis and Management*, 2014, *33* (4), 950–976.

Martinez, S., S. Naudeau, and V. Pereira, "The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique," Technical Report 81532 2012.

Martinez, Sebastian, Sophie Naudeau, and Vitor Azevedo Pereira, "Preschool and Child Development Under Extreme Poverty: Evidence from a Randomized Experiment in Rural Mozambique," World Bank Policy Research Working Paper 8290 2017.

Mendelsohn, Alan L, Leora N Mogilner, Benard P Dreyer, Joel A Forman, Stacey C Weinstein, Monica Broderick, Karyn J Cheng, Tamara Magloire, Taska Moore, and Camille Napier, "The impact of a clinic-based literacy intervention on language development in inner-city preschool children," *Pediatrics*, 2001, *107* (1), 130–134.

Mol, S. E., A. G. Bus, M. T. de Jong, and D. J. H. Smeets, "Added value of dialogic parent-child book readings: A meta-analysis," *Early Education and Development*, 2008, *19* (1), 7–26.

Murray, L., L. De Pascalis, M Tomlinson, Z Vally, H Dadomo, B MacLachlan, C Woodward, and P.J. Cooper, "Randomized controlled trial of a book-sharing intervention in a deprived South African community: effects on carer–infant interactions, and their relation to infant cognitive and socioemotional outcome," *Journal of Child Psychology and Psychiatry*, 2016, *57*, 1370–1379.

Ninio, A., "Joint book reading as a multiple vocabulary acquisition device," *Developmental Psychology*, 1983, *19* (3), 445–451.

Olken, Benjamin A, "Promises and perils of pre-analysis plans," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Opel, A., S. S. Ameer, and F. E. Aboud, "The effect of preschool dialogic reading on vocabulary among rural Bangladeshi children," *International Journal of Educational Research*, 2009, *48* (1), 12–20.

Özler, Berk, Lia C. H. Fernald, Patricia Kariger, Christin McConnell, Michelle Neuman, and Eduardo Fraga, "Combining Pre-School Teacher Training with Parenting Education: A Cluster-Randomized Controlled Trial," *Journal of Development Economics*, 2018, *133*, 448–467.

Paxson, Christina and Norbert Schady, "Cognitive development among young children in Ecuador: the roles of wealth, health, and parenting," *Journal of Human Resources*, 2007, *42* (1), 49–84.

Piper, B. and E. Miksic, "Mother Tongue and Reading: Using Early Grade Reading Assessments to Investigate Language-of-Instruction Policy in East Africa," in A. Gove and A. Wetterberg, eds., *The Early Grade Reading Assessment: Applications and Interventions to Improve Basic Literacy*, Research Triangle Park, NC: RTI Press, 2011.

___ **, S. S. Zuilkowski, D. Kwayumba, and A. Oyanga**, "Examining the secondary effects of mother-tongue literacy instruction in Kenya: Impacts on student learning in English, Kiswahili, and mathematics," *International Journal of Educational Development*, 2018, *59*, 110–127.

Piper, Benjamin, "Kenya Early Grade Reading Assessment Findings Report," published by RTI International 2010.

___ **, Stephanie S Zuilkowski, and Salome Ong'ele**, "Implementing mother tongue instruction in the real world: Results from a medium-scale randomized controlled trial in Kenya," *Comparative Education Review*, 2016, *60* (4), 776–807.

Prado, Elizabeth L, Leila M Larson, Katherine Cox, Kory Bettencourt, Julianne N Kubes, and Anuraj H Shankar, "Do effects of early life interventions on linear growth correspond to effects on neurobehavioural development? A systematic review and meta-analysis," *The Lancet Global Health*, 2019, *7* (10), e1398–e1413.

Reynolds, Sarah A, Chris Andersen, Jere Behrman, Abhijeet Singh, Aryeh D Stein, Liza Benny, Benjamin T Crookston, Santiago Cueto, Kirk Dearden, Andreas Georgiadis, Sonya Krutikova, and Lia C.H. Fernald, "Disparities in children's vocabulary and height in relation to household wealth and parental schooling: a longitudinal study in four low-and middle-income countries," *SSM-Population Health*, 2017, *3*, 767–786.

RTI International, "Early Grade Reading Assessment (EGRA) Toolkit, Second Edition," 2015.

Sabarwal, Shwetlena, David Evans, and Anastasia Marshak, "The Permanent Input Hypothesis: The Case of Textbooks and (No) Student Learning in Sierra Leone," World Bank Policy Research Working Paper 7021 2014.

Simsek, Z. C. and N. I. Erdogan, "Effects of the dialogic and traditional reading techniques on children's language development," *Procedia-Social and Behavioral Sciences*, 2015, *197*, 754–758.

Trudell, B., "Local Community Perspectives and Language of Education in Sub-Saharan African Communities," *International Journal of Educational Development*, 2007, *27* (5), 552–563.

UNICEF, "State of The Worlds Children 2017: Children in a Digital World," Technical Report, New York 2017.

Uwezo, "Are Our Children Learning? The State of Educaiton in Kenya in 2015 and Beyond," Technical Report 2015.

Vally, Zahir, Lynne Murray, Mark Tomlinson, and Peter J Cooper, "The impact of dialogic book-sharing training on infant language and attention: a randomized controlled trial in a deprived South African community," *Journal of Child Psychology and Psychiatry*, 2015, *56* (8), 865–873.

Wasik, B. A. and A. H. Hindman, "Talk alone won't close the 30-million word gap," *Child Development*, 2015, *96* (6), 50–54.

Weber, Ann, Anne Fernald, and Yatma Diop, "When cultural norms discourage talking to babies: effectiveness of a parenting program in rural Senegal," *Child Development*, 2017, *88* (5), 1513–1526.

Weitzman, C. C., L. Roy, T. Walls, and R. Tomlin, "More evidence for reach out and read: a home-based study," *Pediatrics*, 2004, *113* (5), 1248–1253.

Whitehurst, G. J., F. L. Falco, C. J. Lonigan, B. D. Fischel J. E.and DeBaryshe, M. C. Valdez-Menchaca, and M. Caulfield, "Accelerating language development through picture book reading," *Developmental Psychology*, 1988, *24* (4), 552–559.

Wolf, Sharon, Larry Aber, Jere Behrman, and Morgan Peele, "Longitudinal causal impacts of preschool teacher training on Ghanaian children's school readiness: Evidence for persistence and fadeout," *Developmental Science*, 2019, p. e12878.

World Bank, *Ending Learning Poverty: What Will It Take?*, Washington, DC: World Bank, 2019.

Zevenbergen, A. A. and G. J. Whitehurst, "Dialogic reading: A shared picture book reading intervention for preschoolers," in A. van Kleeck, S. A. Stahl, and E. B. Bauer, eds., *On Reading Books to Children: Parents and Teachers*, 2003, pp. 177–200.

Table 1: Summary Statistics on Baseline EMERGE Sample

	Obs.	Mean	S.D.	Median	Min.	Max.
Panel A: Household Characteristics						
Household size	2527	5.88	1.91	6	2	17
Distance to primary school (in meters)	2527	437.30	168.31	453.78	16.45	983.16
Has cement floor	2527	0.16	0.36	0	0	1
Has iron roof	2527	0.97	0.16	1	0	1
Has latrine or toilet	2527	0.77	0.42	1	0	1
Has solar power	2527	0.44	0.50	0	0	1
Connected to power grid	2527	0.08	0.28	0	0	1
Owns a bicycle	2527	0.37	0.48	0	0	1
Owns a car	2527	0.02	0.15	0	0	1
Panel B: Caregiver Characteristics						
Caregiver is child's mother	2527	0.86	0.35	1	0	1
Caregiver is child's father	2527	0.01	0.09	0	0	1
Caregiver is child's grandmother	2527	0.11	0.31	0	0	1
Caregiver illiterate	2518	0.50	0.50	1	0	1
Panel C: Maternal Characteristics						
Child's mother is alive	2527	0.97	0.16	1	0	1
Mother's age	2460	30.67	6.98	30	16	60
Mother is Luo	2527	0.95	0.22	1	0	1
Mother's education in years	2527	7.73	2.40	8	0	13
Panel D: Child Characteristics						
Child is male	2527	0.50	0.50	1	0	1
Child age (in months)	2527	59.26	13.66	60	36	83
Height-for-age z-score	2481	-0.42	1.38	-0.48	-5.81	5.99
Child is enrolled in school	2527	0.86	0.35	1	0	1
Expressive vocabulary (out of 31)	2527	9.48	4.84	9	0	26
Receptive vocabulary in Luo (out of 27)	2527	10.10	5.69	10	0	26
Receptive vocabulary in English (out of 34)	2527	6.62	3.98	6	0	21
Familiarity with letters (out of 9)	2527	2.91	3.17	2	0	9
Familiar word reading in Luo (out of 4)	2527	0.23	0.83	0	0	4
Familiar word reading in English (out of 4)	2527	0.25	0.82	0	0	4
Math skills (out of 6)	2527	2.13	1.89	2	0	6
Fine motor index (out of 6)	2527	4.50	1.64	5	0	6
Panel E: Home Literacy Environment and Parental Investments						
Family Care Indicators score (out of 18)	2527	9.00	3.31	9	1	18
Number of health investments (out of 8)	2527	6.13	1.03	6	0	8

Sample includes data on 2,013 caregivers and 2,527 children aged 36 to 83 months.

Table 2: Baseline Characteristics by Treatment Status: Community-Level Randomization

	Treatment		Control		Difference	
	Mean	S.D.	Mean	S.D.	Diff.	S.E.
Distance to primary school (in meters)	429.53	170.35	445.12	165.93	-13.30	10.30
Has cement floor	0.16	0.37	0.15	0.36	0.01	0.02
Has iron roof	0.99	0.12	0.96	0.19	0.03**	0.01
Has latrine or toilet	0.76	0.42	0.77	0.42	-0.00	0.03
Has solar power	0.42	0.49	0.45	0.50	-0.02	0.03
Connected to power grid	0.10	0.30	0.07	0.25	0.02	0.02
Owns a bicycle	0.36	0.48	0.39	0.49	-0.03	0.03
Owns a car	0.02	0.15	0.02	0.14	0.00	0.01
Household asset index	-0.01	2.07	-0.01	2.01	-0.01	0.10
Caregiver is child's father	0.01	0.09	0.01	0.09	-0.00	0.00
Caregiver is child's grandmother	0.11	0.31	0.11	0.31	-0.00	0.02
Child's mother is alive	0.97	0.16	0.98	0.15	-0.00	0.01
Mother's age	30.85	7.04	30.53	6.91	0.37	0.31
Mother is Luo	0.94	0.24	0.96	0.20	-0.02**	0.01
Receptive vocabulary in English (out of 34)	6.57	3.90	6.68	4.06	-0.14	0.18
Familiarity with letters (out of 9)	2.88	3.13	2.94	3.21	-0.07	0.15
Familiar word reading in Luo (out of 4)	0.20	0.80	0.25	0.86	-0.04	0.03
Familiar word reading in English (out of 4)	0.23	0.80	0.26	0.84	-0.03	0.04
Math skills (out of 6)	2.05	1.87	2.21	1.91	-0.17**	0.08
Fine motor index (out of 6)	4.49	1.63	4.52	1.64	-0.03	0.06
Number of health investments (out of 8)	6.11	1.04	6.16	1.01	-0.05	0.05

Standard deviations in brackets. The Difference column lists the OLS coefficient on the indicator for random assignment to the treatment group from a regression of the covariate on treatment controlling for stratum fixed effects, clustering at the community level; robust standard errors are reported in parentheses. Statistical significance: ***, **, and * indicate significance at the 1, 5, and 10 percent levels, respectively.

Table 3: Midline (2018) Impacts of EMERGE Intervention

Outcome	**Control Mean**	**Coef.**	**S.E.**	**P-Val.**	**95% CI**
Panel A: Books and Comprehension (N=379)					
Any storybooks in home	0.078	0.884	0.019	0.000	[0.844, 0.924]
Number of storybooks in home	0.140	4.588	0.101	0.000	[4.379, 4.798]
Correct comprehension answers (out of 11)	0.404	4.566	0.247	0.000	[4.052, 5.080]
Panel B: Reading Behaviors(N=482)					
Someone read with child (past 3 days)	0.429	0.321	0.040	0.000	[0.238, 0.405]
Caregiver ever read with child	0.409	0.493	0.033	0.000	[0.425, 0.561]
Mother read with child (past week)	0.283	0.364	0.043	0.000	[0.274, 0.454]
Father read with child (past week)	0.113	0.052	0.033	0.138	[-0.018, 0.121]
Sister read with child (past week)	0.192	0.109	0.039	0.010	[0.029, 0.190]
Brother read with child (past week)	0.171	-0.040	0.038	0.306	[-0.120, 0.039]
Grandmother read with child (past week)	0.017	0.056	0.015	0.002	[0.024, 0.088]
Grandfather read with child (past week)	0.004	-0.002	0.004	0.648	[-0.011, 0.007]
Panel C: Observing Caregiver Read with Child (N=364)					
Periods with any interactive activity (out of 20)	13.905	1.451	0.284	0.000	[0.859, 2.042]
Periods with asking child questions (out of 20)	11.503	1.657	0.428	0.001	[0.766, 2.548]
Periods with expanding (out of 20)	0.860	0.590	0.277	0.045	[0.014, 1.165]
Periods with asking child to expand (out of 20)	4.955	2.820	0.614	0.000	[1.543, 4.098]
Periods with answering questions (out of 20)	0.140	0.091	0.048	0.071	[-0.009, 0.191]

Survey conducted for 379 children, each with a distinct caregiver. For most questions asked of caregivers (about up to two children in the home), 482 observations are available. For observation of the caregiver reading with a child, 364 observations are available. All specifications include stratum and age fixed effects. Standard errors clustered at the level of the community.

Table 4: Impacts of EMERGE Intervention in Individually-Randomized Pilot Study (2015)

Outcome	**Coef.**	**S.E.**	**P-Val.**	**95% CI**
Storybook Expressive	0.240	0.113	0.036	[0.016, 0.463]
Non-Storybook Expressive	0.041	0.107	0.703	[-0.171, 0.253]
Luo Receptive Vocabulary	0.022	0.139	0.869	[-0.242, 0.286]
English Receptive Vocabulary	0.147	0.135	0.276	[-0.119, 0.413]
Vocabulary Index	0.113	0.067	0.096	[-0.020, 0.246]

Table 5: Minimum Detectable Effects with and without Baseline Data

				Unadjusted		**Covariate-Adjusted**		
H_0	**Outcome**	N	BL data?	ρ	**MDE**	$\tilde{\sigma}$	$\tilde{\rho}$	**MDE**
1[a]	Storybook Expressive	2,527	Yes	0.076	0.213	0.770	0.026	0.119
1[a]	Non-Storybook Expressive	2,527	Yes	0.050	0.185	0.656	0.010	0.086
1[a]	Luo Receptive	2,527	Yes	0.016	0.140	0.774	0	0.087
1[a]	English Receptive	2,527	Yes	0.026	0.155	0.888	0.005	0.108
1	Vocabulary Index	2,527	Yes	0.060	0.196	0.606	0.015	0.084
2a[b]	Literacy Index (BL only)	2,527	No	0.046	0.180		[f]	
2b[b]	Literacy Index (BL+OS)	4,570	No	0.046	0.165		[f]	
3[c]	Vocabulary Index	1,895	Yes	0.060	0.208	0.606	0.015	0.095
4a[b,c]	Literacy Index (BL only)	1,895	No	0.046	0.193		[f]	
4b[b,c]	Literacy Index (BL+OS)	3,427	No	0.046	0.159		[f]	
7[d]	Vocabulary Index	1,267	Yes	0.478	0.166	0.606	0.026	0.096
8a[b,d]	Literacy Index (BL only)	1,267	No	0.430	0.165		[f]	
8b[b,d]	Literacy Index (BL+OS)	3,424	No	0.430	0.145		[f]	
9	Storybooks in home	2,013	No	0.017	0.150		[g]	
10	Reading Frequency	2,970	Yes[e]	0.034	0.157		[f]	
11	Child Book Familiarity	2,970	No	0.144	0.269		[g]	
12	Caregiver Book Familiarity	2,013	No	0.144	0.274		[g]	
13	Sibling Book Familiarity	2,043	No	0.144	0.274		[g]	
14	Child Stimulation	2,970	Yes	0.094	0.223	0.824	0.010	0.100
15	Demand for Storybooks	2,013	No	0.017	0.150		[g]	
–	Impacts on Older Siblings	2,043	No	$\rho \in [0, 0.1] \Rightarrow \text{MDE} \in [0.125, 0.239]$				
–	Impacts on Caregivers	2,013	No	$\rho \in [0, 0.05] \Rightarrow \text{MDE} \in [0.124, 0.190]$				

[a] We report power calculations for the individual components of the Vocabulary Index for illustrative purposes. [b] In Section 4.3.1 we outline the two potential estimation approaches for testing Hypothesis 2, and the rule that will be used to choose between them once endline outcome data is available. [c] Sample restricted to control group plus households randomly assigned to the Luo books treatment; note that Hypotheses 5, 6a, and 6b would involve sample sizes, sample structures, and power calculations nearly identical to those of Hypotheses 3, 4a, and 4b, but for language (Luo or English). [d] Sample restricted to treatment group to test hypotheses comparing Luo storybooks to English storybooks; clustering at household rather than community level (since storybook language is randomized at the household level within treatment communities). [e] As discussed in Section 4.4.2, we collected baseline data on only one of the five outcomes included in the reading frequency index. [f] For some outcomes (including the Literacy Index and the measures of Reading Frequency, we have some data at baseline that should be predictive: for example, letter and familiar word recognition, and a subset of reading frequency measures. However, this is not a baseline measure of the exact outcome, and we do not have a good way to extrapolate how much σ or ρ will change with these baseline variables included as controls. We expect some improvement from the unadjusted measure, but we include this not as a flag of some uncertainty about additional power. [g] For other measures (Book Familiarity, Storybooks in home, Demand for Storybooks), we are less confident that any baseline variable will substantially improve power.

A Appendix: Sample Images from Child Assessments

Figure A1: Sample Storybook Comprehension Questions

Panel A.

What 3 things did Caro forget at home today?

Panel B.

What is Meg holding?

Panel C.

What do you think Meg and Ben are looking at?

Figure A2: Sample Expressive Vocabulary Stimuli

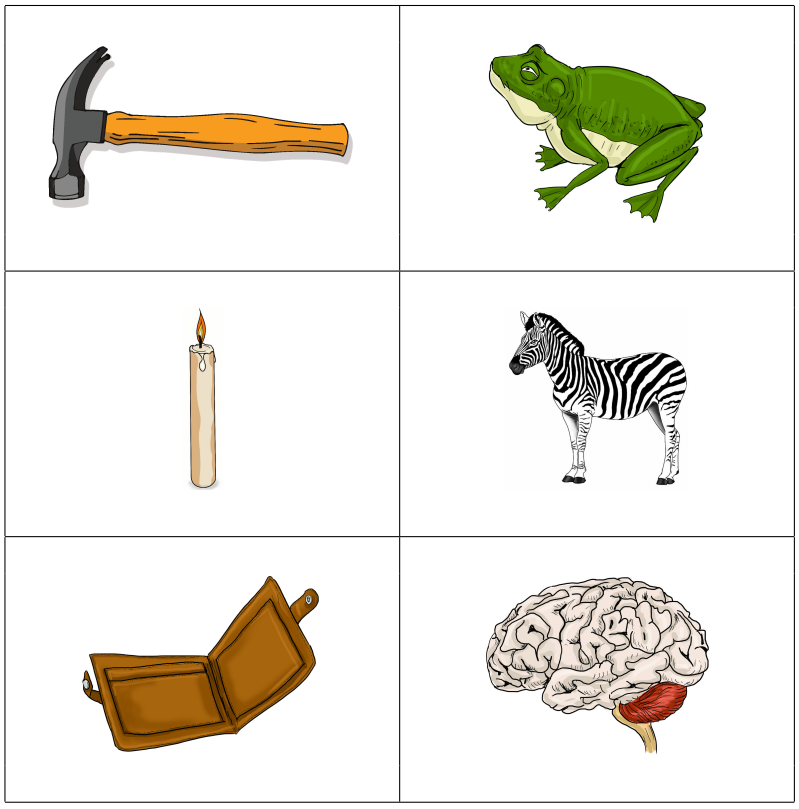
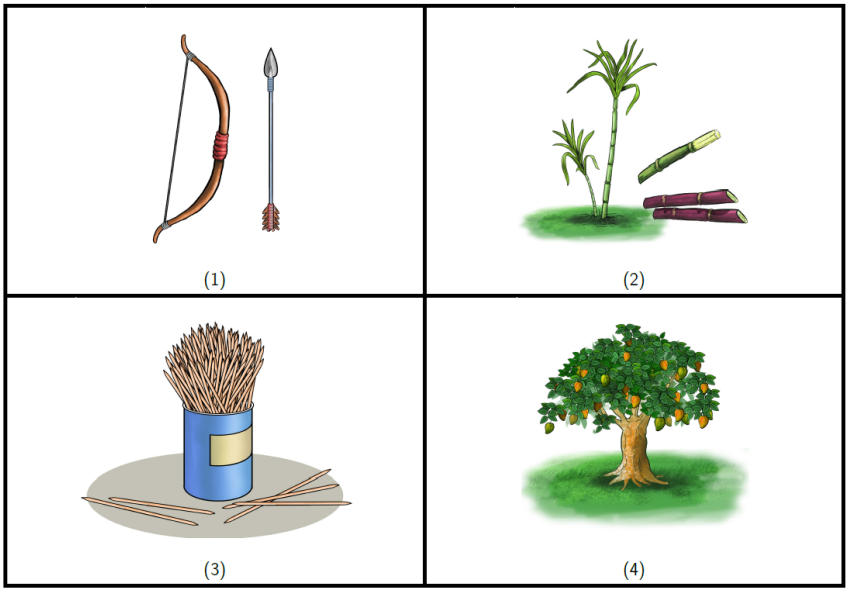
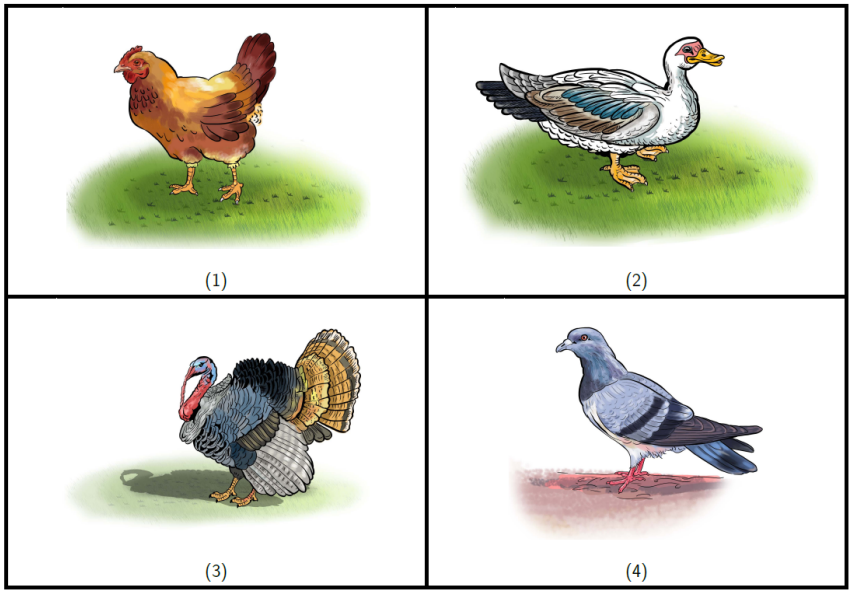


Figure A3: Sample Receptive Vocabulary Stimuli





It is not appropriate to describe contributions before a paper is complete, in this case, for the Stage 1 registered report.