WPS7817

Policy Research Working Paper 7817

Combining Preschool Teacher Training with Parenting Education

A Cluster-Randomized Controlled Trial

Berk Özler
Lia C. H. Fernald
Patricia Kariger
Christin McConnell
Michelle Neuman
Eduardo Fraga

WORLD BANK GROUP
Development Research Group
Poverty and Inequality Team
September 2016

Policy Research Working Paper 7817

Abstract

This paper evaluates a government program in Malawi, which aimed to improve quality at community-based childcare centers and complemented these efforts with a group-based parenting support program. Children in the integrated intervention arm (teacher training and parenting) had significantly higher scores in measures of language and socio-emotional development than children in centers receiving teacher training alone at the 18-month follow-up. However, the study finds no effects on child assessments at the 36-month follow-up. Significant improvements at the centers relating to classroom organization and teacher behavior in the teacher-training only arm did not translate into improvements in child outcomes at either follow-up. The findings suggest that, in resource-poor settings with informal preschools, programs that integrate parenting support within preschools may be more effective than programs that simply improve classroom quality.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at bozler@worldbank.org.

Combining Preschool Teacher Training with Parenting Education:

A Cluster-Randomized Controlled Trial

Berk Özler, Development Research Group, The World Bank*

Lia C. H. Fernald, School of Public Health, University of California, Berkeley

Patricia Kariger, School of Public Health, University of California, Berkeley

Christin McConnell, Education Global Practice, The World Bank

Michelle Neuman, Graduate School of Education, University of Pennsylvania

Eduardo Fraga, Department of Economics, Yale University

Keywords: early childhood development, parenting education, preschool teacher training

JEL Codes: J24, J13, I20

* Corresponding author: bozler@worldbank.org. The study obtained approvals from the University of California at Berkeley (CPHS Protocol Number: 2011-07-3464) and from the Malawian National Commission for Science and Technology (Reference Number: RTT/2/20).

1. Introduction

Early life investments can be beneficial for children both for short- and longer-term outcomes, such as better labor market returns (e.g. (Gertler et al., 2014), and programs targeting disadvantaged children can be socially efficient (Elango et al., 2015). Returns to investments in early childhood are higher than investments made later in life because beneficiaries have a longer time to reap the rewards, and because early childhood is a sensitive period during which adverse exposures as well as positive interventions can have the greatest effects on an individual's developmental trajectory (Heckman & Mosso, 2014). Furthermore, early investments in human capital have dynamic complementarities, such that "learning begets learning" (Carneiro & Heckman, 2003).[1]

Lack of adequate preparation for primary school through pre-primary education is one of the key risk factors for poor performance in primary school (Behrman et al., 2006).[2] Thus, a popular approach to trying to improve outcomes in children has to do with increasing enrollment in preschool programs, and/or trying to improve the quality of existing programs.[3] Children in low-resource settings are less likely to attend school, and they are not likely to learn when they are in the school setting – partly because they are unprepared for school when they get there.

---

[1] Early child development (ECD) investments can take many forms, including promotion of good health and nutrition, support for safe and stimulating environments, protection from risks such as violence or abandonment, parenting support and early learning experiences, media, preschools, and community groups. Systematic reviews of ECD interventions in LMICs have shown success, particularly when interventions are high quality, targeted to the most vulnerable children and integrated with other services (Engle et al., 2011).

[2] Pre-primary (or preschool) programs generally refer to an organized learning group that meets at least two hours per week and can be categorized as formal (institutionalized, intentional and planned through public organizations) or informal (not institutionalized, less structured, less organized) (UNESCO, 2011).

[3] Enrollment in preschool programs has increased substantially in LMICs over the past several decades (Behrman et al., 2013), but is still far from optimal. UNESCO's Global Monitoring Report of pre-primary enrollment rates in 2012 showed coverage ranging from 19% (86%) for low-income (high-income) countries (UNESCO, 2015).

Most studies comparing attendees of preschools of any type (e.g. formal or informal) with non-attendees have found higher scores on some measure of child development, such as literacy, vocabulary, math, and quantitative reasoning, teacher assessments at the end of the year, and/or on subsequent school performance (for reviews, see (Engle et al., 2011; Rao et al., 2015)).[4] Beyond these comparisons, there is also a growing literature and emphasis on quality, and what variables are most important for developing and supporting a young child's abilities (Behrman et al., 2013; Britto et al., 2014).

The evaluation of Chile's *Un Buen Comienzo* (A Good Start) is the first large-scale, randomized study of an effort to improve the quality of preschool education in South America (Yoshikawa et al., 2015). It is a two-year program that provides teacher training and professional development to prekindergarten and kindergarten teachers. After exposure to the teacher-training intervention, many classroom characteristics and teacher behaviors showed significant improvements. There were no effects on children's language or literacy skills, however. The authors interpret these findings as a consequence of the low intensity of, and insufficient exposure of the children to, the training program.

Another approach to promoting early child development revolves around support of parents.[5] Most existing studies of home-visiting studies in low- or middle-income countries (LMICs) have been smaller efficacy trials, though some recent papers have examined programs at scale. For example, a home visiting program in Pakistan utilized community health workers and demonstrated improved child development outcomes (Yousafzai et al., 2014). A recent scaled-up program in the Caribbean delivered parenting support messages within primary care clinics and showed benefits to child development (Chang

[4] For example, preschool attendance compared with non-attendance has been associated with better cognitive performance among preschool children in Mozambique (Martinez et al., 2012); reduced dropout and grade repetition among children in Uruguay (Berlinski et al., 2008); better school performance among third graders in Argentina (Berlinski et al., 2009); and improved 4th grade math scores in a national sample in Brazil (Rodrigues et al., 2010).

[5] Meta-analyses of parenting and home visiting programs from high- and low- income countries have found that the most effective parenting programs included systematic training methods, a structured, evidence-based curriculum built on a strong, theory-driven approach (Engle et al., 2011; Segal et al., 2012).

et al., 2015). A study set in Mexico utilized the existing structure of the country's conditional cash transfer program (*Prospera*) to deliver group-based parenting support and showed positive effects on child development (Fernald et al., in press), as did a similar study in Colombia, which used a home-visiting approach (Attanasio et al., 2014). An explanation for the Mexican findings was the increase in the number of play activities that parents engaged in with children, which led to improved child development outcomes (Knauer et al., 2016). There were also modest, positive effects of an adult literacy and parental participation program in India on outcomes in children aged 5-8 years old (Banerji et al., 2015).

In this study, using a cluster-randomized controlled trial, we test the effectiveness of teacher training at informal schools in a resource-poor setting on early childhood development and primary school readiness to assess whether such school-based interventions are more effective when combined with group-based parenting training. Because the newly trained teachers and mentors in the community deliver the group-based parenting training, it is easily and cheaply scalable. Given the consistent effectiveness of parenting support in the promotion of early child development, and the widespread use of informal preschools within (LMICs) (Garcia et al., 2008), our findings have the potential for broad policy relevance within a context of extreme poverty and limited government resources.

Our study focuses on Community-Based Childcare Centers (CBCCs) in Malawi, which are widespread in the country and estimated to serve 580,000 children in approximately 5,000 communities (Drouin & Heymann, 2010). In its wish to support these CBCCs, instead of setting up a parallel formal preschool sector, the government decided to improve the supply of play and learning materials in these centers and strengthen the capacity of teachers through additional training and mentoring to support children's early development and learning. To evaluate the marginal effectiveness of the teacher-training program over and above the provision of play and learning materials, we designed an experiment where the comparison group received only a standard kit of supplies from UNICEF, while a second arm also received teacher training and mentoring.

To test two additional variations to this model of training volunteer teachers in existing informal preschools, we added a third arm within which the trained teachers were assigned to receive a small

monthly stipend during the first school year following the intervention. This was intended to increase retention and motivation among these otherwise unpaid workers. In the fourth and final arm of the trial, we complemented the school-based teacher-training program with a 12-module, group-based, parenting education program for the primary caregivers of the children enrolled at the CBCC. This program used the teachers and their mentors as parenting education facilitators and focused on increasing parental engagement by teaching them specific tools for stimulating their children's cognitive development (e.g. by reading or playing with them), and promoting their health (e.g. hand washing, nutrition, etc.) at home. This model has the advantage of being cheaper than stand-alone home-visiting programs, which made it more suitable for a poor country like Malawi; there is the additional advantage of using the newly trained teachers and mentors that makes this model scalable.

We find that primary child outcomes improved at the 18-month follow-up (when the average child in our study sample was 5.5 years old), but only in the treatment group that received the integrated intervention – with teacher training and parenting education. In this group, children had significantly higher scores in an assessment of language skills and they exhibited more prosocial behaviors when compared with both the control group and the teacher training only group. The gains at the child level from the added parenting education were accompanied by substantial improvements in family care indicators, e.g. how many times a day their primary caregivers read to their children or played with them.

Teacher training alone (or with monthly stipends for retention) did not improve children's outcomes, despite significant improvements relating to the classroom environment and teacher behaviors. Furthermore, a rich battery of child assessments, conducted 36 months after baseline, showed no treatment effects among the 6-8 year-old children in any treatment arm, indicating a substantial fade-out of program impacts in the integrated intervention arm.

Our paper makes several contributions to the literature on early childhood investments. First, the evidence base on the effects of providing training to preschool teachers is scant, particularly in poor and informal settings. In contrast to much of Latin America, we examine the effects of teacher training in a context where the workforce is untrained and unpaid, which is not uncommon in Sub-Saharan Africa

(Garcia et al., 2008). CBCCs in Malawi rely on volunteer workers with minimal training and low levels of education.

Second, our findings echo those of Chile's *Un Buen Comienzo*, which also found that improvements in classroom quality did not translate into improvements in child-level outcomes at the end of the two-year teacher training intervention. Being assigned to higher quality classrooms in kindergarten has been recently shown to modestly increase math, language, and executive function test scores among children (Araujo et al., 2016). Yet, the study by (Yoshikawa et al., 2015) in Chile and our study in Malawi highlight the difficulty of converting program-induced improvements in classroom quality into better child outcomes.

Third, our trial has incorporated parenting support into the context of a preschool-based quality improvement intervention. The two main existing approaches to early childhood investments – preschool quality improvements and parenting support – have not previously been tested together. The fact that the group-based parenting support sessions were delivered by the newly trained preschool teachers made this integrated intervention easily and cheaply scalable, but unfortunately ruled out a more classical two-by-two factorial design. As such, we are unable to speak to the cost-effectiveness of parenting support alone, but we found promising evidence that this approach can improve child outcomes over and above teacher training – at least in the short run.

Finally, the trial design also allows us to identify the causal effect of exogenously improving classroom and parenting quality on child development outcomes. As teacher training strongly improved observed classroom quality over and above the provision of play and learning materials to the control group, we can identify the causal relationship between classroom quality and child outcomes using an instrumental variables approach. Similarly, as the integrated intervention substantially improved parenting quality over and above teacher training, we can speak to the effect of improved parenting quality on child development. Our analysis suggests that, in this context, the effect of classroom quality improvements is negligible while those of parenting quality are significant and large. It is worth noting that the integrated

intervention is the only arm in which numeracy, literacy, and problem-solving activities increased both in the classroom and at home.

The paper proceeds as follows. Section 2 describes the study design. Section 3 describes our estimation strategy. Section 4 presents our findings, while Section 5 provides concluding remarks and ideas for future research.

2. Study Design

Study Setting

The Government of Malawi (termed Government hereafter) and development partners have supported a model of community-initiated and -owned centers, known as Community-based Childcare Centers (CBCCs), particularly since the 1990s. CBCCs were meant to promote holistic child development by providing safe, stimulating environments, access to health and nutrition services, and capacity building for teachers (Munthali et al., 2008; Munthali et al., 2014).

However, perhaps because CBCCs in Malawi are designed to be self-sustaining – owned, managed, and operated by the communities themselves – the quality of the facilities and services provided remain quite poor, particularly in comparison to richer countries. School facilities range from permanent structures such as private homes, churches, old shops, and NGO-sponsored community centers to thatch structures and decrepit shelters (Munthali et al., 2014). CBCCs generally have a covered indoor space with burnt-brick walls and concrete or mud floors, as well as a cooking area where food is prepared. Most CBCCs lack basic play and learning materials, including a lack of books for children. Almost all CBCCs depend on small financial or in-kind contributions from community members and parents to cover costs.

Initially intended for custodial care, CBCCs in our study districts operate for a few hours each weekday morning and are run by teachers, who are typically untrained, unpaid, and largely female. Less than half of the teachers have received the government-developed 14-day training module and nearly a third have worked at their respective center for less than a year (Fisher et al., 2009; Ministry, 2010). Overall, the education level of teachers at CBCCs is low: baseline data indicate that a third of them lacked

a primary school leaving certificate (PSLC), which is obtained by passing an exam at the end of Grade 8 in Malawi. Lack of compensation and incentives hamper the recruitment and retention of teachers.

Study design and sample selection

Our study is a cluster-randomized controlled trial with four arms (Figure 1): a control group and three treatment arms, all of which are described in detail under the *Interventions* sub-section below. Four study districts (Balaka, Dedza, Nhkata Bay, and Thyolo) representing all three regions of Malawi (Southern, Central, and Northern) were chosen by the Ministry of Gender, Children, Disability, and Social Welfare (termed Ministry hereafter). A full listing of all CBCCs eligible to receive the intervention produced only 199 CBCCs in these four districts (Neuman et al., 2014). Sample size calculations for a multi-site, cluster-randomized trial showed that the detectable difference between any two study arms for a standardized child assessment with an intra-cluster correlation of 0.1 would be approximately 0.25 standard deviations with 95% confidence and 80% power if we sampled 12 children per CBCC with 50 CBCCs allocated to each arm. Therefore, all 199 CBCCs were selected for inclusion in the study.

Children were randomly selected (blocked by age and sex) from the group of children attending the CBCCs on the day the baseline data collection teams visited the school.[6] The study sample includes 2,120 children (an average of 10.7 children per CBCC), aged 36-61 months at baseline, from the study centers. The study also enrolled 2,009 primary caregivers of the sampled children, who provided consent for the sampled children to be included in the study and completed the primary caregiver questionnaire;

[6] Specifically, in each CBCC, all of the children who were in attendance during the visit of the baseline data collection team were split into four groups with the help of the teacher: three-year-old boys, three-year-old girls, four-year-old boys, and four-year-old girls. The total number of children in each group was recorded to enable the construction of sampling weights used in our analysis. Each group then formed a circle with the help of the teacher and a child was selected with the help of a random number generator. Then, every third child in the circle was selected until three (or all, whichever was higher) children were selected from each group. This procedure produced a median (mean) of 12 (10.7) children per center, for a total sample size of 2,120 children in 199 centers.

some of the primary caregivers have more than one child randomly selected for participation in the study.[7]

After baseline data collection was completed, random assignment procedures for the allocation of CBCCs to the four study arms were conducted in each district separately. To boost statistical power and ensure a balance of important baseline characteristics, a "block randomization" procedure was used. Centers were grouped based on mean height-for-age (HAZ) and Peabody Picture Vocabulary Test (PPVT – a measure of receptive vocabulary) z-scores, both of which were collected at baseline. The Ministry held a public lottery at each district capital, where a representative from each center was asked to draw a colored dot from an envelope to determine that center's treatment status.[8] The representatives of the centers assigned to the same treatment group in each district were then invited to a private information

[7] Consent forms (in English and Chichewa) are available from the authors upon request.

[8] The public lotteries took place in early 2012: January in Balaka, February in Thyolo, and March in Dedza and Nkhata Bay. We calculated mean HAZ and PPVT scores for each CBCC (the CBCC means were calculated using inverse probability sampling weights to make the means representative of the population of eligible children in the study sample). We first sorted the CBCCs by mean HAZ to form several groups of CBCCs. Then, we sorted the CBCCs within each of these blocks by their mean PPVT scores to form bins of four CBCCs (when the number of sampled CBCCs in a district was not divisible by four, there was a leftover group with less than four centers). These data were then provided to our counterpart in the Ministry, who wrote the names of the schools in each bin on a piece of paper and taped it onto an envelope that contained four different colored dots. Representatives of each CBCC drew a dot from their assigned envelopes during the public lottery to determine the center's treatment status. In three of the four study districts, this procedure was followed perfectly. In the fourth district, our document with the list of schools assigned to each bin did not reach Ministry officials in time for the public lottery; in this case, they created their own bins and conducted the rest of the public lottery as planned. As described in the next section, we use the *actual* bins used for the public lotteries as controls in our estimation of program effects, three of which are the same as the *intended* bins by the research team, while one is different.

session immediately following the lottery, where a Ministry official provided them with details about the study and their assigned intervention. Hence, the centers were not blinded to their treatment status and could later learn about treatment variations received by other centers in their district.

Interventions

The Protecting Early Childhood Development Project (PECD), co-designed by the Ministry, World Bank, and academic partners from UC Berkeley to ensure technical, scientific, and policy relevance, was designed to test strategies to improve the quality and stability of existing CBCCs. The interventions focused on: (a) improving the play and learning resources in CBCCs; (b) strengthening the capacity of teachers to support children's early development and learning; (c) teaching parents about how to support development and learning activities in the home. Under PECD, the Government implemented the following interventions – in partnership with Save the Children and UNICEF:

T1. *Comparison Group: Provision of play and learning materials*

To address the basic developmental and learning needs of children, each center in the study received a kit of basic play and learning materials and supplies procured by UNICEF. The contents of the kit were developed by the Ministry and included items such as books, displays, balls, paint, chalk, blocks, puzzles, first aid kit, and kitchen utensils.

T2. *T1 + Training and mentoring of teachers*

To improve the knowledge, skills, and practices of teachers in the 150 CBCCs assigned to the three treatment groups, the project tested an enhanced teacher-training package. The teacher-training component followed a cascade model, where national ECD specialists trained regional trainers, who then trained the teachers at CBCCs, and aimed to build early childhood development (ECD) capacity at the national, district, and community levels.[9] Each CBCC in a treatment group nominated two teachers to

[9] Before launching the training, a series of training guides and manuals were developed and validated by national ECD experts in Malawi. Nine National Core ECD trainers participated in a five-day orientation course facilitated by representatives from the South Africa-based Early Learning Resource Unit, a World Bank ECD Consultant from

participate in the training program. Training candidates were required to hold a Junior Certificate of Education (JCE), which requires passing an exam at the end of Form 2 (or Grade 10) in Malawi. If none of the active teachers at the CBCC met this minimum qualification, it was recommended that the CBCC seek and nominate an eligible volunteer from the same village. However, in practice, candidates with only a PSLC were nominated by the centers and were accepted for training.

The teacher-training program consisted of five weeks of residential training divided into two two-week sessions and a final one-week session. Between these sessions, the teachers went back to their CBCCs for a few months to practice their newly acquired skills in the classroom with support from supervisors and mentors. During these periods, the teachers in training held briefings with the untrained teachers at the CBCC. 310 teachers from 150 treatment CBCCs (out of a total of 468 teachers at these CBCCs) completed all three phases of training. The program covered the following modules: child development; play and early learning materials and equipment; learning through play; planning and organizing the learning environment; child health and care; child hygiene and environmental care; child nutrition and care; child rights and their welfare; care and development of children with special needs; early childhood care, management, and partnership.

Mentors and supervisors were trained for three days after the first training session. Mentors were teachers selected based on exceptional performance, reputation in the community, and commitment to their work. 37 mentors provided guidance and support to teachers in their respective districts through weekly visits. Each mentor was assigned to four CBCCs translating to a ratio of one mentor to eight teachers in training. Supervisors were Child Protection Workers (CPWs) and government-employed

---

Kenya, and Ministry officials. The National ECD trainers had responsibility for training, supporting, and supervising regional trainers and serving as an advisory group for the PECD training program throughout the intervention period. Thirty-five regional trainers were competitively selected by the Ministry to train CBCC teachers, mentors, supervisors, Center Management Committees (CMCs) and parenting facilitators. All regional trainers held at least a Malawi School Certificate of Education (MSCE), had completed basic ECD training, and were active in the ECD field. They participated in a three-week residential training facilitated by the National ECD trainers.

social workers, who already worked within the area of the study CBCCs. These 19 supervisors were each responsible for supervising eight CBCCs on average. National and district officials, along with Save the Children staff, also conducted monitoring visits. Finally, to strengthen the capacity of Center Management Committees (CMC) in carrying out their responsibilities to manage the centers, the project included a five-day training conducted within the communities, which reached 1,499 committee members from 150 CBCCs.

T3. *T2 + Teacher incentives*

In each of the 49 CBCCs assigned to this group, the teachers who successfully completed the teacher-training program were given a small monthly stipend (MK 2,000) for a period of seven months to increase retention and perhaps improve motivation. Unfortunately, the intervention period coincided with an economic crisis in Malawi, which saw the value of this payment depreciate from US$12 to $6 per month from April to November 2012. Save the Children personnel administered the payments and ensured that the trained teachers received the incentives. In several communities, CMCs raised some funds to make comparable payments to the teachers who did not receive the enhanced teacher training.

T4. *T2 + Parenting education*

Primary caregivers of children attending the 51 CBCCs assigned to this arm participated in group sessions that provided information and demonstrated practical activities that they could replicate at home. Parent educators – CPWs, trained teachers, and mentors, who received three days of training for this task – facilitated the sessions. The initial implementation of the parenting education arm deviated from the original project design in that the 12 parenting sessions were first conducted in 12 consecutive days instead of 12 days spread out over six weeks. To rectify this implementation error, starting approximately one month later, the facilitators held six additional “refresher” sessions of two hours each for six weeks. In between these weekly sessions, parents and guardians were encouraged to practice with their children at home what they had learned.

The sessions covered the following topics: introduction to child development; physical development; mathematical and critical thinking; general knowledge and scientific thinking; language

development; literacy; social and emotional development; spiritual and moral development; supporting children's approaches to learning; children's health and safety; children's nutrition and food safety. Participants used items from the kits provided to the CBCCs by UNICEF along with locally available materials, such as leaves, stones, soil, feathers, bean bags, charcoal, drums, etc.

Data Sources

We have so far conducted three rounds of data collection: baseline (Round 1), 18-month follow-up (Round 2), and 36-month follow-up (Round 3). In what follows we use 18-month (36-month) follow-up and Round 2 (3) interchangeably.

Child Measures

A comprehensive battery of child development measures was used to assess language, fine motor, executive function (attention, inhibition, working memory), problem solving, social/emotional and numeracy/math skills. These measures cover abilities that typically begin to emerge and progress early in life; are encouraged through commonly recommended preschool practices; and are believed to be important for primary school success (Copple & Bredekamp, 2009; Duncan et al., 2007; Sabol & Pianta, 2012). All selected assessments, which are described in detail in the Appendix, had demonstrated reliability and/or validity in either Malawi or other Sub-Saharan countries. Each test was translated and adapted as necessary for use in the present study. At the 36-month follow-up (Round 3), some scales were dropped because they no longer showed good variability in performance (i.e., were too easy) or because the children had aged out (e.g. Malawi Developmental Assessment Tool, or MDAT), while other tests indicative of expanding capacities (e.g. Early Grade Mathematics Assessment, or EGMA; and Kaufman Assessment Battery for Children, or KABC) were added.[10] Appendix Table 1 shows the schedule of child

[10] We also considered the Chichewa version of the Early Grade Reading Assessment (EGRA), but found most children piloted could not recognize letters, parts of words, or words. Our pilot was similar to findings from a 2010 National report showing 76.5% of children starting Standard 2 could not name a single letter, and 92% could not read familiar words (Mejia, 2010)

assessments at each round of data collection. The rules used to score each assessment, and to aggregate subscales into indices are also described in great detail in the Appendix.

Anthropometric measurements were made at baseline to (i) assess balance across groups, (ii) control for any direct or indirect influences growth faltering (specifically stunting or chronic malnutrition) might have on child assessments, and (iii) assess heterogeneity of treatment effects. Child height and weight were measured according to the nearest 0.1 cm and 0.1 kg, respectively, following established guidelines (Cogill, 2003). Height-for-age (HAZ), weight-for-height (WHZ), and weight-for-age Z-scores (WAZ) were then calculated using the 2006 World Health Organization (WHO) growth standards (WHO Multicentre Growth Reference Study Group, 2006).

Primary Caregiver Measures

In addition to gathering data on household characteristics, we collected data on the primary caregiver's health and the home environment. At baseline and the 18-month follow-up (Round 2), the primary caregiver's mental health status was assessed. At all rounds, the provision of household stimulation for the child's learning and development, as well as the use of positive disciplinary techniques was measured. A standardized parenting quality index combining these three scales was created for Rounds 1 and 2. The scales administered to each caregiver are described in detail in the Appendix; scales that were child specific (e.g. the Parenting Stress Index) were administered once for each child.

CBCC Measures

Extensive information on the characteristics of the CBCC, staff, and quality of staff-child interactions was gathered at baseline and both follow-up rounds. The CBCC questionnaire and observation measure were adapted from the *La Escala de Evaluación de la Calidad Educativa de Centros de Educación Preescolares* (ECCP) (Martínez et al., 2004) from Mexico and a preschool quality tool used in Cambodia (Rao et al., 2012). Key information from the questionnaire included teacher characteristics, such as education, experience, and training.

Classroom observations were conducted while the CBCC was operating to provide an objective account of classroom organization, activities and teacher-child interactions. To complete the observations,

pairs of trained fieldworkers arrived unannounced just as the CBCC was opening, and observed normal center activities for one hour.[11] The enumerators rated the classroom environment across a variety of indicators that included, for example, teachers' styles of teaching various concepts, encouragement of child participation in learning, time spent reading, time spent engaged with children (either individually or in groups), response to children's needs, disciplinary strategies, use of small and large groups, and interactions that promote children's social development. The rules used to score the classroom observations, and to aggregate subscales into an overall classroom quality index are described in great detail in the Appendix.

3. Estimation Strategy

We take advantage of the randomized allocation of the interventions at the cluster (CBCC) level to construct estimation models for causal identification. As described in the data collection and instrument sub-sections above, baseline data were collected when the children were 3-5 years old. Our primary outcomes are child assessments conducted 18 and 36 months after baseline – when the children are 4.5-6.5 and 6-8 years old, respectively. We analyze these outcomes at the individual level. To estimate intention-to-treat (ITT) effects of each intervention on child outcomes, we employ a regression model of the following form for each round of follow-up data collection:

$$Y_{ij} = \alpha + \gamma^2 T_j^2 + \gamma^3 T_j^3 + \gamma^4 T_j^4 + \beta X_{ij} + \varepsilon_{ij} \quad (1),$$

where Y_{ij} is an outcome variable for child *i* in CBCC *j*, T_j^2, T_j^3, and $T_j^4$are binary indicators for CBCC-level interventions T2 through T4, and X_{ij} is a vector of baseline characteristics. The standard errors ε_{ij}, clustered at the CBCC level, account for both the design effect of the cluster-level treatment and

[11] The duration of classroom observations was raised to two hours in Round 3 due to the possibility that treatment could affect the timing of classroom activities rather than their quantity or quality.

heteroskedasticity inherent in the regression model. Age- and sex-specific sampling weights are used to make the results representative of the target population of children in the study centers.[12]

For each measure of our primary outcomes in rounds 2 and 3, we estimate two versions of the model in equation (1). In the "unadjusted" regressions, we only include indicators for the strata used to perform block randomization – i.e. the "district x bin" fixed effects, where bins refer to the groups of four CBCCs on the envelopes used during the public lotteries that were described in detail above (Bruhn & McKenzie, 2009). In our "adjusted" regressions, we add indicator variables for child age in months and the baseline (lagged) value of the child development measure to the X_{ij} vector. These variables were chosen because they are strongly predictive of performance at follow-up and, as a result, improve the precision of the impact estimates. We prefer this analysis of covariance specification to a difference-in-difference estimation because of the large gains in power (McKenzie, 2012).

After analyzing the relative effectiveness of each of the interventions on child outcomes, we examine effects of the interventions on secondary outcomes reported by the primary caregiver – such as family care indicators, positive parenting practices, and the parenting stress index – in the same way we analyze the child outcomes. As caregivers with more than one child in the study sample were asked about each child separately, these regressions are also at the child level. We examine these variables to understand the mechanisms that may explain the intervention effects: for example, there may have been effects of the intervention on the behavior of the child's primary caregiver, especially in T4, which combined efforts to improve basic quality at the CBCCs with parenting support.

Finally, we examine changes at the CBCC level, also as an attempt to understand the mediating factors in the observed impacts on child outcomes, using the following approach:

$$Y_j = \alpha + \gamma^2 T_j^2 + \gamma^3 T_j^3 + \gamma^4 T_j^4 + \beta X_j + \varepsilon_j \qquad (2),$$

[12] Regression models estimating program impacts without the use of sampling weights (not shown here) produce estimates that are very similar to the weighted estimates reported in this paper.

where Y_j is an outcome variable for CBCC j, and X_j is baseline (lagged) value of the outcome. As in the earlier analysis at the child level, all regressions include "district x bin" fixed effects.

4. Findings

In this section, we first discuss baseline balance and attrition to establish the internal validity of the impact estimates. We then present program effects at the 18- and 36-month follow-ups. Within each round, we first present primary outcomes at the child level followed by secondary outcomes at the principal caregiver and CBCC levels. We conclude our presentation of findings with an examination of heterogeneity of impacts by mother's education and child height.

Baseline balance and attrition

Table 1 presents baseline balance for child-level characteristics, where we see consistent balance across all treatment arms for four child assessments conducted at baseline (Leiter Sustained Attention, Peabody Picture Vocabulary Test, and Malawi Developmental Assessment Tool language and fine motor/perception subscales), as well as age in months, gender, and HAZ. Chi-squared tests for joint orthogonality never produce a p-value below 0.297 for any of the six pairwise comparisons of treatment arms.[13] Appendix Tables 2 and 3 present caregiver- and CBCC-level characteristics and show good balance across treatment arms as well. CBCCs in T4 display significantly higher classroom quality scores at baseline than the control group and T2, but as we will see in Tables 6 and 11, this baseline characteristic is not predictive of classroom quality in the two follow-up rounds and is therefore unlikely to cause bias in our impact estimates.

In Table 2, we present, for each round, the overall level of attrition, whether it varies by treatment arms, and whether attrition is differentially correlated with baseline characteristics. Attrition of children from the sample is small at 0.062 in the control group during the 18-month follow-up and 0.046 during

[13] A chi-squared test for the joint significance of 21 coefficient estimates (seven variables for each of the three interventions) after a multinomial logistic regression produces a p-value of 0.84.

the 36-month follow-up.[14] There is no sign of differential attrition for any of the treatment groups. Joint F-tests of interaction terms suggest no differences in the types of children who are lost to follow-up in either round. We conclude that the loss of children from the study sample to either follow-up is not consequential to causal inference in this experiment. Appendix Table 4 shows that attrition levels among the primary caregivers are naturally similar to those for the children (at 0.063 and 0.045, respectively in rounds 2 and 3) and caregivers in T2 are slightly less likely to be lost to follow-up in round 2 (p-value=0.067). While there is no sign that the attrition in this group is correlated with baseline characteristics of the caregiver, there are some random differences in the baseline characteristics of primary caregivers lost to Round 2 follow-up in T3 and T4. There are no differences in levels or types of attrition among primary caregivers in round 3.

Given their informal nature, the CBCCs in Malawi are likely to have been closed and reopened over the course of the study period. Appendix Table 5 shows that 6.1% of the CBCCs in the control group were not operational in Round 2 – with this level of attrition being similar across all treatment arms.[15] In Round 3, a larger (10.2% of the control group) and different set of centers was closed (meaning that many centers that were closed in Round 2 had reopened and were surveyed); centers in T2 were slightly more likely to be closed at Round 3. Joint F-tests of baseline controls and interactions with treatment indicators in either round indicate no signs of systematic attrition in the control group or differential attrition with respect to baseline characteristics except for T3 at the 18-month follow-up.

[14] At the 18-month follow-up, we randomly sampled 42 children out of 127, who were not originally found but were categorized as "trackable," for further tracking. We successfully found and assessed 35 of these 42 children. Our attrition analysis for Round 2, in columns 1 & 2 in Table 2, assigns higher weights to these 42 children and excludes the 85 children who were "randomized out," which explain the sample size of 2,035 rather than 2,120.

[15] Enumerators who had an unsuccessful initial visit to a CBCC were instructed to revisit it two more times before reporting it as closed.

In summary, the data show a good balance of baseline characteristics across treatment arms and little problem with attrition in either round of data collection – particularly for the primary outcomes at the child level. We now proceed to presenting estimates of ITT effects at the 18-month follow-up.

18-month impacts

Primary outcomes (child level)

At 18-months, we conducted child assessments using the MDAT, described in the Appendix, which evaluates language and fine motor/perception skills in children and was specifically created for use in rural Malawi.[16] Table 3 presents program effects on the overall score, as well as the two subscales (Language and Fine Motor/Perception Skills). Columns (1) and (2) present impacts by treatment arm on the total MDAT score: there were no effects in T2 or T3, but there was a significant effect of T4 compared to the control group (0.13 SD, p-value=0.06; column (2)). In columns (3) and (4), we see that this improvement in T4 is due to the intervention's effect on language skills, which is 0.19 SD (p-value=0.010). This improvement is significantly higher than that in T2 (T2-T4=0.22, p-value=0.001), indicating that marginal value of the parenting education is substantial over and above teacher training at the CBCC level. However, it is not significantly larger than the effect in T3, where the teachers were provided small incentives for retention and motivation (T3-T4=0.1; p-value=0.17). Columns (5) and (6) show no effects of T4 on Fine Motor/Perception Skills in comparison to the control group. T4 still performs significantly better than both T2 and T3, but this is due to a negative impact of these two interventions (only significant at the 90% level).

To assess program impacts on positive and negative behaviors among the target population of children, we also administered the Strengths and Difficulties Questionnaire (SDQs, described in detail in

[16] Detailed information on all assessments – what they measure, how they were scored, and how they were aggregated into larger indices – can be found in the appendix. All impact findings are robust to scoring child assessments using item response theory (IRT) rather than the simpler standard scoring rules that were used in our analysis (findings not shown here).

the Appendix), which is reported by the primary caregiver of each child. Table 4 shows the effects on the child's behavioral problems (inverted *total difficulties score*, in which a higher score indicates a lower level of behavioral problems) and the child's positive behaviors (*prosocial score*, in which a higher score indicates more prosocial behaviors). We detect no significant effects on the total difficulties score in any of the three treatment groups compared with the control group, but children in T3 and T4 exhibit significantly lower levels of behavioral problems than children in T2 (columns (1) and (2)). Estimates of program effects presented in Panel A of Appendix Table 6 by the four subscales of the total difficulties score (Emotion, Conduct, Hyperactivity, and Peer Problems) suggest that benefits of T3 and T4 over T2 are driven mainly by improvements in conduct and hyperactivity. In columns (3) and (4) of Table 4, we see that children in T4 are reported to be substantially more prosocial (0.25 SD; p-value=0.001) compared with the control group – with the difference between this group and the other two intervention arms also being meaningful and statistically significant.

In summary, at the 18-month follow-up, children assigned to T4 showed significant but moderate improvements in language skills, as well as increased prosocial behaviors, when compared with the control group.[17] All of these beneficial effects are significantly higher in T4 than in T2, which suggests that the parenting education was an effective addition to the teacher training provided at the CBCC level. In the next sub-section, we investigate secondary outcomes at the caregiver and CBCC levels, some of which might have mediated the positive effects of T4 on child outcomes.

Secondary outcomes (primary caregiver level)

During the survey with the primary caregiver for each child, we asked about activities that adults do with children to encourage learning, such as reading books to them, telling them stories, singing, playing with the children or helping them learn letters, numbers, colors, shapes, identify objects, etc. The

[17] Tables 3 & 4 present 12 t-tests of treatment effects (four primary outcomes in three treatment arms). Our finding of statistically significant effects of T4 on language skills and prosocial behaviors remains intact when we control for "false discovery rates" by calculating q-values (Benjamini & Hochberg, 1995) as described in (Anderson, 2008).

list of activities was adjusted across rounds of data collection to be age-appropriate and we constructed a standardized index of indicators at each round of data collection. We created three summary indicators: the Stimulation Index (the activities with the child), the Positive Practices Index (positively dealing with behavioral issues of the child), and the Parenting Stress Index (caregiver's stress related to raising the child) – all of which are described in the Appendix. These three indices are further aggregated into an overall parenting quality index using inverse covariance weighting and standardized in each round.[18]

Table 5 presents program impacts on parenting quality. The group assigned to parenting training (T4) experienced a large and highly significant improvement in parenting quality (0.26 SD, p-value<0.01; columns (1) and (2)). There were no impacts in T2 or T3, which is expected since T4 is the only intervention that had a component targeted at the primary caregiver. The impact on parenting quality in T4 is also significantly higher than these two other treatment arms. In columns (3) through (8), we present program impacts on each of the scales in the parenting quality index: we note that the effect in T4 is largely due to a large and significant increase on the number of activities that adults in the households were reported to have done with their children (0.29 SD, p-value<0.001). Appendix Table 7 presents impacts on each item in the Stimulation Index and shows that adults in T4 households are significantly more likely than every other group to report reading books, playing and chatting with their children, and helping them learn letters, numbers, colors, and shapes at the 18-month follow-up (Panel A). These effects are consistent with program effects in T4 on children's language assessments presented in Table 3.

Secondary outcomes (CBCC level)

As described earlier, during each round of data collection, we had two enumerators show up unannounced at each CBCC to observe the classroom and then jointly fill out a detailed questionnaire to describe the nature and the quality of various activities. These activities were grouped into *routine and*

[18] Inverse covariance weighting (ICW) assumes one latent trait of interest that underlies the set of items and tries to construct an optimal weighted average based on that assumption (Samii, 2016). We followed (Anderson, 2008) and (Casey et al., 2012) to construct the ICW weights. More details are provided in the Appendix.

structure; *supervision of children*; *teacher engagement with the children*; *dealing with children's behaviors*; *communication with the children*; *numeracy, literacy, and problem solving*; *fine and gross motor activities*; and *miscellaneous activities* (including music and movement, science and nature, and spirituality). To avoid assigning ad hoc, such as equal, weights to each of these subscales to form an overall index (or to each question within a subscale to form a subscale index), we again used inverse covariance weighting to create a standardized overall index.

Table 6 presents program impacts on changes in classroom practices (columns (1) and (2)). Sizeable and significant program impacts are apparent in all treatment arms, consistent with the fact that each intervention contained intensive teacher training and some mentoring. However, the effects are substantially larger in T3 and T4 than in T2 (p-values 0.024 and 0.112, respectively), suggesting that the small incentives provided (T3) and the act of providing parenting education within the community (T4) may have improved classroom practices over and above teacher training and mentoring alone.

Appendix Table 8 provides an explanation for this heterogeneity in treatment effects by showing impacts on components of the classroom observation tool. In columns (1) and (2) of Panel A, we show program impacts on two components of the classroom quality index, which we constructed ourselves by combining pre-defined sub-scales: one component consists of items under *routine and structure*; *supervision of children*; *teacher engagement with the children*; *dealing with children's behaviors*; and *communication with the children*; while the other one contains *numeracy, literacy, and problem solving*; *fine and gross motor activities*; and *miscellaneous activities* (including music and movement, and science and nature). We see that while all treatment centers show large improvements in the former component, increased activities in *numeracy, literacy, and problem solving* and *fine and gross motor activities* are observed only in T3 and T4, but not in T2. We also used principal components analysis to allow the data to identify the latent orthogonal factors underlying the classroom observation index: we present impacts on the first two components in columns (3) and (4). Interestingly, the data also yield similar factor loadings: the first component contains items from *supervision* and *engagement* subscales (e.g. children were less likely to be left unsupervised, teachers were more likely to be sitting rather than standing during

engagement with children, etc.), while the second factor is mostly comprised of items from the *numeracy, literacy, and problem solving* subscales (e.g. teaching numbers and the alphabet, reading with children, identifying shapes and colors, etc.). As a result, the findings are qualitatively the same as those in columns (1) and (2).

Columns (3) and (4) in Table 6 report changes in enrollment. Interestingly, enrollment increased significantly in treatment schools (T2 and T4), which may indicate that parents value improvements in school quality over and above the basic learning and play materials received by the control schools (T1). The effects range from a statistically insignificant 10% increase over the control group in T3 to significant increases of 17% and 21% in T2 and T4, respectively – with none of the increases in the three treatment arms significantly different than each other.

If the parents think that quality is improved at treated centers, it may be the case that these schools have more qualified and/or better-trained staff. Column (5) shows program impacts on the number of teachers trained by the program. There are no trained teachers in control CBCCs, compared with approximately 1.5 trained teachers across treatment groups, with no significant differences between any of the treated CBCCs. Given that each school nominated two teachers for training under the program, which is confirmed by administrative records of the training sessions provided by Save the Children, this indicates that roughly one out of four trained teachers was no longer teaching by the 18-month follow-up. Despite this loss of approximately 25% of the trained teachers, treatment schools had, as intended, a significantly higher number of trained teachers than the control group. Contrary to our original hypotheses, the small incentives offered to trained teachers in T3 did not improve their retention.[19]

[19] In Appendix Table 12, we further examine changes in the characteristics of the teachers. The share of teachers who have a primary school leaving certificate (PSLC) is significantly higher in treatment schools, particularly in T3 – the intervention that provided small incentives to trained teachers. While educated teachers are more likely to have left the center by Round 2 in the control group, those without PSLCs are more likely to have left the CBCC in treatment centers. Furthermore, new arrivals in Round 2 are significantly more likely to have PSLCs in the treatment

In summary, we find significant effects on primary child outcomes in the short-run, which are supported by impacts on secondary outcomes at the primary caregiver and CBCC levels. Interestingly, substantial improvements in classroom practices in T2 and, in particular, T3 were insufficient to improve child outcomes. Only in T4, where improved parenting at home by the primary caregiver reinforced such improvements at the CBCC level, we find effects on child outcomes.

Causal effects of classroom and parenting quality on child assessments

In Tables 5 and 6, we showed strong program effects on parenting and classroom quality, respectively. In particular, teacher training combined with parenting education improved self-reported parenting quality substantially over and above teacher training alone (F-test for T4-T2=0 is approximately 24), while teacher training with incentives improved classroom practices over and above simply receiving teaching and play materials (F-test for T3-T1=0 is approximately 70). Each of these strong impact estimates can serve as the first-stage in an instrumental variables (IV) approach and allow us to answer the following question: what is the causal effect of improving parenting (classroom) quality on child assessments?[20] Given that we carefully constructed comprehensive indices of classroom observations and parenting quality, it is reasonable to assume that the exclusion restriction holds in each case.[21]

---

group. Treatment schools also had a younger teaching staff at Round 2. Hence, the program led to a younger and more educated group of teachers in all treatment arms.

[20] We have also conducted the same IV exercise using T2 and the control group. While the first stage is weaker (F-test for T2-T1=0 is approximately 9), the finding of a null effect of classroom quality on child assessments is robust to this alternative specification.

[21] Teachers trained under T2 and T3 do not interact with the children outside of the CBCC. Our data show no effect of these interventions on the number of times teachers meet with parents, the median for which is zero in the control group. Hence, it is reasonable to assume that the effect of T3 on child assessments must only be mediated by what happens in the classroom. With parenting quality, it is possible that parenting education affects aspects of parenting (such as self-esteem or confidence) that are not included in our parenting quality index. Or, the act of providing parenting education to primary caregivers may improve teaching quality in the classroom over and above receiving

Table 7 presents our findings. In panel A, using only T2 and T4, we see that a one standard deviation increase in parenting quality leads to large and significant increases in all child assessments. The IV coefficient estimates range between 0.42 SD (MDAT Fine Motor skills) and 0.71 (MDAT Language skills). In Panel B, this time using only T3 (which provide us with a stronger first-stage regression than T2) and the control group, we find that increases in our classroom observation index do not cause any significant changes in language or fine motor skills, but may moderately lower children's behavioral problems. Our findings suggest that while both classroom practices and parenting quality can be successfully manipulated through (modest) interventions like the ones studied here, only the latter consistently caused significant improvements in children's language, cognitive, and socio-emotional skills in this context.

36-month impacts

Primary outcomes (child level)

At the 36-month follow-up, with the average child in our study sample at seven years of age, we conducted a richer set of assessments to gauge program impacts on primary school readiness. The tests we selected evaluate word comprehension (PPVT), memory and problem solving skills (KABC), knowledge of numbers and basic mathematics skills (EGMA), and maintaining attention and accuracy during a task (Leiter Sustained Attention). Hence, they should provide an indication of increased task performance if the program had lasting effects.

Table 8 presents our findings. There are no effects on any of the four domains assessed in any of the three treatment arms and no significant differences between any two arms. Appendix Table 9 presents impacts on the subscales of Kaufman and EGMA: of the 18 adjusted coefficient estimates (eight

---

teacher training only, for which we presented evidence in Table 6. However, a similar-sized improvement in classroom quality in T3 did not lead to any improvements in child assessments, making it less likely that the effect of T4 (over and above T2) is through this alternative channel of classroom quality. In any case, our findings in this short sub-section should be interpreted as being only suggestive of a causal relationship between each of these two important concepts and children's developmental outcomes.

outcomes – six subscales plus PPVT and Leiter Sustained Attention – by three interventions), only two are statistically significant – one of which is positive (T3 on KABC triangles subscale) and the other one negative (T2 on EGMA number recognition subscale). The null findings are not due to a lack of power: the standard errors in Table 8 indicate that we would have been able to detect effect sizes of 0.13-0.18 SD on all child development assessments.

The readers will note the lack of overlap for child assessments used in Rounds 2 and 3. While the Kaufman and EGMA modules were added in Round 3, the MDAT was discontinued.[22] As mentioned earlier, MDAT was discontinued in Round 3 simply because the children aged out of it. EGMA (early grade mathematics assessment) was added because almost all the children in our study sample were of primary school age by Round 3 and, hence, expected to have started recognizing numbers, discriminating quantities, and conducting simple addition.

The different child assessments used at the 18- and 36-month follow-ups make it harder to interpret our findings: did the program effects, especially on language skills in T4, really fade out or are the changes in assessments responsible for our findings? Appendix Table 10 first presents the cross-sectional correlation coefficients between all child assessments administered at baseline: we can see that MDAT language skills and PPVT are highly correlated with each other (0.57, p-value<0.01), much higher than the correlation coefficients between PPVT and any other assessment administered at baseline. Appendix Table 10 also shows that the baseline MDAT language score is as good a predictor of the PPVT score at the 36-month follow-up as the lagged baseline value of PPVT itself. Furthermore, the correlation coefficient between the 18-month MDAT language score and the 36-month PPVT score is 0.30 (p-value<0.01). Given the high cross-sectional correlation between the two language assessments administered at baseline plus the fact that baseline MDAT and PPVT scores are equally strong predictors

[22] Due to budgetary reasons, we had to make a choice between administering the MDAT or the PPVT in Round 2 and opted for the former, which was specifically designed and culturally appropriate for Malawi, worked well at baseline, and was highly correlated with the PPVT.

of a child's PPVT score at the 36-month follow-up, we find it more likely that the fadeout of T4 effects on language skills is real rather than an artifact of our assessment tools.

Table 9 shows that there are no effects on either the (inverted) SDQ total difficulties score or the prosocial behaviors subscale – although children in T4 displayed more behavioral problems than those in T2 and T3. Again, Panel B in Appendix Table 6 presents treatment effects for the four subscales of the SDQ *total difficulties score*. The disappearance of the strong 18-month impacts of the combined parenting intervention on both the total difficulties score and the prosocial index at the 36-month follow-up reinforces the idea that the fadeout of program impacts on child outcomes is real. In Appendix Table 11, we present further analysis of impacts at the primary school level, which almost all children in the study sample attended by Round 3. This final investigation also reveals no impacts on either the children's school attendance, grade progression, and repetition, or their parents' likelihood of discussing their children's progress or behaviors with their primary school teachers.

Secondary outcomes (primary caregiver level)

Table 10 presents program effects on parenting quality in Round 3. As in Round 2, we created a Stimulation Index (activities that adults in the household do with the children in our study sample to encourage learning), and Positive Practices Index (positive methods caregivers use to deal with children's problem behaviors), but we did not administer the Parenting Stress Index in Round 3. In columns (1) and (2), caregivers in T4 still report significantly higher levels of activities with their children, but the size of this standardized effect, at 0.16 SD is approximately half of what it was in Round 2 (0.29 SD). This effect on the Stimulation Index is still significantly larger than the effect in T2, meaning that the parenting education had a lasting effect on the behavior of primary caregivers. However, Panel B of Appendix Table 7 shows that while some effects on reading with children remain, the effects on helping them with literacy and numeracy disappeared over time. This fadeout is consistent with our finding of null effects on language (PPVT) and mathematics (EGMA) at the 36-month follow-up, despite significant improvements in language skills 18 months earlier. As in Round 2, there are no effects on the Positive Practices Index (columns (3) and (4)).

Secondary outcomes (CBCC level)

The findings in Round 3 at the CBCC level, presented in Table 11, are consistent with the effects at the child and caregiver levels: classroom observations indicate slightly better-quality classrooms among the treatment groups in general, but the effect sizes are much smaller than Round 2 and none of the effects is significant at the 90% level of confidence (columns (1) and (2)). Panel B of Appendix Table 8 shows that the consistent and large program effects on the sub-components of classroom quality at the 18-month follow-up have similarly faded out. Effects on enrollment similarly decayed and are not statistically significant at the 36-month follow-up (columns (3) and (4)). Finally, while the treatment schools lost some more trained teachers to departures, each treatment group still has approximately 1.25 more teachers trained by the program 36 months after baseline (column (5)) and a higher percentage of the teaching staff at treatment schools have PSLCs (Appendix Table 12).

In summary, the effects observed at the 18-month follow-up disappeared completely for the primary outcomes and faded out significantly for the secondary outcomes by the 36-month follow-up. We found no effects on any child assessments or the caregiver-reported SDQs, and many of the secondary outcomes at the primary caregiver and CBCC levels were smaller or had also vanished.

Heterogeneity of impacts

Low height-for-age, or stunting, is recognized as a risk factor for child development and international development organizations have recently been putting more emphasis on investing in programs that can close the gap in cognitive outcomes between stunted and non-stunted children. Also, as Banerji et al. (2015) discuss, parenting education programs for parents, who themselves have had very little formal education or are illiterate, may not be effective. As we could not experiment with combining parenting support with an adult literacy or a nutrition intervention, we briefly examine the heterogeneity of effects at the 18-month follow-up by mother's education and child height at baseline.

Regression specifications presented Table 12 include centered covariates for child's age in months, child's height-for-age z-score (HAZ), and whether the primary caregiver has a primary school leaving certificate (PSLC), the lagged value of the outcome variable, and their interactions with each

treatment arm.[23] ITT effects reported in the first three rows are very similar to the effects reported in Tables 3-5, which implies that the program effects reported earlier are robust to the inclusion of more baseline controls, most of which are strongly predictive of the outcomes, particularly child assessments.

We find suggestive evidence that the interventions are more effective in raising MDAT scores for children who are shorter for their age: all nine interaction terms between treatment groups and HAZ reported in columns (1)-(3) are negative, with two reaching statistical significance: a one standard deviation increase in baseline HAZ is associated with a decrease in the impact of T4 (T2) on MDAT Language (Total) score by 0.12 (0.13) standard deviations.

When we examine the moderating effects of the primary caregiver's education, we see that the entire effect of T4 (teacher training plus parenting education) on the MDAT language skills at the 18-month follow-up is driven by the effect on children with primary caregivers who have a PSLC at baseline. While the effect of T4 on language is small and insignificant for the approximately 80% of the children whose primary caregivers do not have a PSLC, it is large and significant among the minority of children whose mothers have completed primary school (0.51 SD; p-value-0.06). Consistent with Banerji et al. (2015), the parenting intervention and primary caregiver's education appear to complement each other and, worse, the parenting intervention seems completely ineffective for children with uneducated parents.

5. Concluding Discussion

Our key findings were that children benefited in the short term (18-month follow-up, when they were 4.5 to 6.5 years old) from the integrated intervention arm, which combined teacher training with

[23] Following Lin, Green, and Coppock (2015) and Imbens and Rubin (2015), we include centered covariates and their interactions with treatment indicators in this fully adjusted specification. Centered covariates are simply linearly transformed versions of the baseline covariates by subtracting the sample mean. Then, the coefficient for the (uninteracted) treatment indicator still gives us the average treatment effect for the entire sample (Lin, 2013). In this specification, we replace the month of age dummies with a discrete variable for age in months. For brevity, we only report the interaction terms for HAZ and PSLC, along with the ITT effects at the top.

group-based parenting education. Children in this group had significantly higher scores in assessments of language and socio-emotional development when compared with the comparison group, and, more importantly, when compared with the teacher training only group. The benefits for children in this group might be explained by family care indicators, which demonstrated improvements in terms of the frequency with which parents reported playing or reading with their children. Although all mothers in the integrated intervention arm reported improved family care indicators, beneficial effects on child development were concentrated among children of educated mothers. These findings suggest that mothers with more education may have been better able to absorb, respond to, and act on the information provided in the parenting training sessions than those with less education. It is also possible that more educated mothers were more likely to bring their children to the parenting sessions, who then benefitted directly from participation, which is a mechanism suggested previously by other studies (Banerji et al., 2015).

There were significant improvements at the CBCC level in terms of classroom resources, student management, and teacher responsiveness in all treatment arms, though classroom activities relating to numeracy/literacy and fine/gross motor skills only improved in T3 and T4. In spite of these improvements to the classroom environment, there were no benefits of teacher training alone (or with stipends for retention) in terms of child outcomes. This finding was surprising given that teacher training is often a component of preschool quality improvement (Behrman et al., 2013). Interestingly, however, these findings are similar to the Chilean study, in which a two-year teacher training had significant effects on classroom characteristics and teacher behaviors but no effects on children's language or literacy skills (Yoshikawa et al., 2015). The most successful arm of our study (T4) may have reinforced messages at home that had been presented in school, suggesting a potential pathway by which benefits to cognition and language could have occurred in the short run. In fact, T4 is the only arm in which numeracy, literacy, and problem-solving activities increased both in the classroom and at home, which might explain why we see significant moderate effects on language skills. To avoid the fade-out in this group, perhaps the intensity of the combined intervention should have been higher, or regular refreshers should have supplemented the existing program across early childhood.

A lingering question that emerges from our findings is why classroom improvements did not translate into improvements in child development outcomes in this study. In the Chilean study mentioned above, the authors speculate that there were no child-level effects due to the low intensity of the training program, and insufficient exposure of the children to the program, both of which could also be true for us. Another explanation could be that our study was not sufficiently powered to detect findings with a small magnitude. For example, in Ecuador where children were assigned to higher quality classrooms in kindergarten and showed increases in math, language, and executive function test scores, the reported effect sizes were 0.1 SD (Araujo et al., 2016). Finally, the lack of an improvement in classroom quality for activities related to *numeracy, literacy, and problem solving* and *fine and gross motor activities* may partly explain the lack of impacts on child assessments in language and fine motor skills in T2.

The ITT effects in the parenting arm at the 18-month follow-up were only slightly smaller than those found in the study of a home-visiting program combined with a CCT program in Colombia (0.26 SD) (Attanasio et al., 2014) but much smaller than the scaled-up, home-visiting program in Pakistan (0.60 SD for cognition) (Yousafzai et al., 2014). It is not surprising that our effect sizes are smaller because the parenting intervention was group-based and low-intensity; it should be noted, however, that the effects we report were quite large among children with educated primary caregivers (greater than 0.5 SD for MDAT language skills). A meta-analysis of home-visiting programs in the United States, found a lower average effect on cognitive outcomes of 0.18 SD (Sweet & Appelbaum, 2004). The intensive Jamaican home-visiting efficacy study found a very large effect size of 0.88 SD (Grantham-McGregor et al., 1991), but effect sizes from other efficacy studies of home-visiting programs in low- and middle-income countries have been much smaller (Grantham-McGregor et al., 2014). Finally, our effect sizes are larger than what was reported for older children in India who were exposed to a scaled-up, parenting literacy and learning

support intervention (0.05 SD) (Banerji et al., 2015) or for children in Ecuador who were randomly assigned to kindergarten teachers (0.10 SD) (Araujo et al., 2016).[24]

At the 36-month follow-up (when the children were 6-8 years old), using a complex battery of child assessments, we found no sustained treatment effects on child development outcomes. Thus, the program did not result in higher levels of school readiness for primary school, despite the early indications of improvements in child development indicators, parental involvement, and improvements to the classroom environment. The reasons for the fade-out of effects may have been that once the children graduated from the CBCCs, they were then absorbed into a primary school system with low quality that could not provide an appropriate or beneficial learning environment for the children to build sustainably on what they had learned in the CBCC.

The primary weakness of our approach has to do with the quality of implementation of the program. Although our approach was grounded in the real-life conditions of rural Malawi, the success of the intervention relied on a volunteer workforce with low levels of education and formal training, which may have had consequences in terms of intervention effectiveness. A second weakness is that we are not able to identify the effects or cost-effectiveness of parenting support alone, since the group-based parenting support sessions were delivered by the newly trained preschool teachers, relying on the teacher training intervention. While this approach made the intervention easier to administer and more likely to be scalable in this resource-constrained environment, it ruled out isolating the direct effect of parenting support alone. A third weakness is that we had a short intervention period that lacked intensity, and outcomes may have been better if regular refreshers supplemented the interventions across early childhood in addition to the short intervention period. Fourth, other than the SDQ, we have no child

[24] Our assessments are comparable with other studies, such as the study of the Ecuadorian children (Araujo et al., 2016), which also used the TVIP, in addition to tests of vocabulary and comprehension from the Woodcock-Johnson battery of child assessment tests; our tests are also comparable to the Chilean study, which used subtests of the Woodcock-Munoz, and other letter-word recognition tasks (Yoshikawa et al., 2015).

assessments that were administered during all three rounds of data collection. However, as discussed in detail in the previous section, we feel that the MDAT Language Skills subscale is highly correlated with and predictive of PPVT scores, which leads us to believe that the fadeout in child assessment gains between the two follow-up rounds of data collection is real rather than an artifact of our assessment schedule. The fadeout in the SDQ scores, which were consistently measured over time throughout the study, supports this hypothesis. Finally, the family care indicators are self-reported measures of household stimulation, which can be subject to social desirability bias.

Our study also has several strengths. First, we used a wide range of tests that covered the key domains of child development: language, cognition, behavioral development and school achievement; the tests were well adapted and extensively piloted before use. Second, our study rigorously tested several intervention approaches, all of which represented small changes to the status quo and are directly scalable within the context of informally administered, community-based preschool systems. Third, PECD targeted children aged 3-5, which is a particularly vulnerable period for development during which children have great plasticity and biological receptivity to interventions: our approach was novel because of the use of a parenting support intervention in this age group, which is usually targeted solely with schooling-based approaches. Finally, taking advantage of our field experiment with multiple treatment arms, we were able to investigate causal effects of parenting and classroom quality using an instrumental variables approach. While these exploratory findings should be considered suggestive, they are consistent with recent work that suggests that improved parenting behaviors (such as reading to children) and home environment are significant mediators of improved child outcomes (Knauer et al., 2016).

The Sustainable Development Goals call for all children to “have access to quality early childhood development, care, and pre-primary education so that they are ready for primary education” by 2030 (UNICEF, 2015). Achievement of the goal requires greater coordination of early child development programming with the broader educational infrastructure, with attention to quality of services. Our results suggest that there can be significant benefits to child development from group-based parenting support in the context of an informal preschool setting, but that the early benefits faded over time. Although the

approach reported here is promising because of its potential for scalability, future interventions will have to be strengthened in order to demonstrate sustained outcomes for children.

References

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

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

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *Institute for the Study of Labor,* Available at http://ftp.iza.org/dp9796.pdf.

Attanasio, O. P., et al. (2014). Using the infrastructure of a conditional cash transfer program to deliver a scalable integrated early child development program in Colombia: cluster randomized controlled trial. *BMJ,* 349, g5785.

Banerji, R., Berry, J., & Shotland, M. (2015). The impact of mother literacy and participation programs: Evidence from a randomized evaluation in India. . *Unpublished manuscript (https://sites.google.com/site/econjimberry/research)*.

Behrman, J. R., Engle, P., & Fernald, L. C. H. (2013). Preschool Programs in Developing Countries. Education Policy in Developing Countries pp. 65-79). Chicago: Chicago University Press.

Behrman, J. R., et al. (2006). What Determines Adult Skills? Impacts of Pre-School, School-Years, and Post-School Experiences in Guatemala. Philadelphia, PA: University of Pennsylvania, mimeo, Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=947480.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate. *Journal of the Royal Statistical Society: Series B,* 57(1), 289-300.

Berlinski, S., Galiani, S., & Gertler, P. (2009). The effect of pre-primary education on primary school performance. *Journal of Public Economics,* 93(1-2), 219-234.

Berlinski, S., Galiani, S., & Manacorda, M. (2008). Giving children a better start: Preschool attendance and school-age profiles. *Journal of Public Economics,* 92(5-6), 1416-1440.

Britto, P. R., et al. (2014). Strengthening systems for integrated early childhood development services: a cross-national analysis of governance. *Ann N Y Acad Sci,* 1308, 245-255.

Bruhn, M., & McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics,* 1(4), 200-232.

Carneiro, P., & Heckman, J. (2003). Human Capital Policy. In J. Heckman & A. Krueger (Eds.), Inequality in America: What role for human capital policies? Cambridge: MIT Press.

Casey, K., Glennerster, R., & Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. *The Quarterly Journal of Economics,* 127(4), 1755-1812.

Chang, S. M., et al. (2015). Integrating a Parenting Intervention With Routine Primary Health Care: A Cluster Randomized Trial. *Pediatrics,* 136(2), 272-280.

Cogill, B. (2003). Anthropometrics Indicators Measurement Guide (http://www.fantaproject.org/downloads/pdfs/anthro_2003.pdf), Accessed July 10, 2012. Washington, D.C.: Food and Nutrition Technical Assistance Project, Academy for Educational Development.

Copple, C., & Bredekamp, S. (2009). Developmentally appropriate practice in early childhood programs serving children from birth through age 8. Washington D.C.: Developmentally appropriate practice in early childhood programs serving children from birth through age 8.

Drouin, O., & Heymann, J. (2010). Scaling up and sustaining community-based care for preschool and school-age children -- successes and challenges in Malawi. *Vulnerable Child Youth Stud,* 5(S1), 31-39.

Duncan, G. J., et al. (2007). School readiness and later achievement. *Developmental Psychology,* 43(6), 1428-1446.

Elango, S., Garcia, J. L., Heckman, J. J., & Hojman, A. (2015). Early Chilhood Education. *NBER Working Paper Series,* Working Paper 21766.

Engle, P. E., et al. (2011). Strategies for reducing inequalities and improving developmental outcomes for young children in low and middle income countries. *The Lancet,* 378(9799), 1339-1353.

Fernald, L. C. H., et al. (in press). Promoting Child Development through Group-Based Parent Support within a Cash Transfer Program: Experimental Effects on Children's Outcomes *Developmental Psychology*.

Fisher, W., Kholowa, F., Chibwana, K., & Silo, L. (2009). Success against the odds, a positive deviance study of community based care centres in Malawi. Lilongwe: UNICEF and Government of Malawi.

Garcia, M., Pence, A., & Evans, J. L. (Eds.) (2008). Africa's Future, Africa's Challenge: Early Childhood Care and Development in Sub-Saharan Africa. Washington: World Bank.

Gertler, P., et al. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science,* 344(6187), 998-1001.

Grantham-McGregor, S. M., Fernald, L. C., Kagawa, R. M., & Walker, S. (2014). Effects of integrated child development and nutrition interventions on child development and nutritional status. *Ann N Y Acad Sci,* 1308, 11-32.

Grantham-McGregor, S. M., Powell, C. A., Walker, S. P., & Himes, J. H. (1991). Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study. *Lancet,* 338(8758), 1-5.

Heckman, J. J., & Mosso, S. (2014). The Economics of Human Development and Social Mobility. *Annu Rev Econom,* 6, 689-733.

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY: Cambridge University Press.

Knauer, H. A., et al. (2016). Pathways to improved development for children living in poverty: A randomized effectiveness trial in rural Mexico. *International Journal of Behavioral Development.*

Lin, W. (2013). Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique. *The Annals of Applied Statistics,* 7(1), 295-318.

Lin, W., Green, D. P., & Coppock, A. (2015). Standard operating procedures for Don Green's lab at Columbia. Available at: https://github.com/acoppock/Green-Lab-SOP/blob/master/Green_Lab_SOP.pdf.

Martínez, J. F., Myers, R., & Linares, M. (2004). ¿Todos los pollos son amarillos?: En búsqueda de la calidad educativa en centros preescolares. [Are all chickens yellow? An investigation into the quality of preschool centers.]. Mexico: Dirección General de Investigación Educativa. Secretaría de Educación Pública.

Martinez, S., Naudeau, S., & Pereira, V. (2012). The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique. Available at http://www.3ieimpact.org/media/filer/2013/04/11/3ie_mozambique_ie001.pdf (Accessed August 9, 2013)

McKenzie, D. (2012). Beyond Baseline and Follow-up: The Case for More T in Experiments. *Journal of Development Economics*, 99(2), 210-221.

Mejia, J. (2010). Early Grade Reading Assessment: National Baseline Report.

Ministry (2010). Annual Report for Early Childhood Development. Ministry of Gender, Children, Disability and Social Welfare (MoGCDSW).

Munthali, A., Mvula, P., & Silo, L. (2008). Community-based childcare centres in Malawi: a national inventory. . University of Malawi, Zomba: Centre for Social Research.

Munthali, A. C., Mvula, P. M., & Silo, L. (2014). Early childhood development: the role of community based childcare centres in Malawi. . *SpringerPlus*, 3, 305.

Neuman, M. J., McConnell, C., & Kholowa, F. (2014). From Early Childhood Development Policy to Sustainability: The Fragility of Community-Based Childcare Services in Malawi. *International Journal of Early Childhood*, 46(1), 81-99.

Rao, N., et al. (2012). Is something better than nothing? An evaluation of early childhood programs in Cambodia. *Child Dev*, 83(3), 864-876.

Rao, N., et al. (2015). Early childhood development and cognitive development in developing countries: a rigorous literature review. *Department for International Development, University of Hong Kong. Available at: http://eppi.ioe.ac.uk/cms/LinkClick.aspx?fileticket=9reL_ORWZmI%3d&tabid=3465*.

Rodrigues, C. G., Pinto, C. X. C., & Santos, D. D. (2010). The impact of daycare attendance on math test scores for a cohort of 4th graders in Brazil. Report to the Inter-American Development Bank.

Sabol, T. J., & Pianta, R. C. (2012). Patterns of school readiness forecast achievement and socioemotional development at the end of elementary school. *Child Development,* 83(1), 282-299.

Samii, C. (2016). Inverse covariance weighting versus factor analysis, http://cyrussamii.com/?p=2177. Accessed on August 23, 2016.

Segal, L., Opie, R., & Dalziel, K. (2012). Theory! The Missing Link in Understanding the Performance of Neonate/Infant Home-Visiting Programs to Prevent Child Maltreatment: A Systematic Review. *Milbank Quarterly,* 90(1), 47-106.

Sweet, M. A., & Appelbaum, M. I. (2004). Is Home Visiting an Effective Strategy? A Meta-Analytic Review of Home Visiting Programs for Families With Young Children. *Child Development,* 75(5), 1435-1456.

UNESCO. (2011). International Standard Classification of Education. Available at http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf. .

UNESCO. (2015). EFA Global Monitoring Report 2015.

UNICEF (2015). Transforming our world: the 2030 Agenda for Sustainable Development. Available at: https://sustainabledevelopment.un.org/post2015/transformingourworld.

World Health Organization Multicentre Growth Reference Study Group (2006). WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatr Suppl,* 450, 76-85.

Yoshikawa, H., et al. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. *Dev Psychol,* 51(3), 309-322.

Yousafzai, A. K., Rasheed, M. A., Rizvi, A., Armstrong, R., & Bhutta, Z. A. (2014). Effect of integrated responsive stimulation and nutrition interventions in the Lady Health Worker programme in Pakistan on child development, growth, and health outcomes: a cluster-randomised factorial effectiveness trial. *Lancet,* 384(9950), 1282-1293.

FIGURE 1: Trial Profile

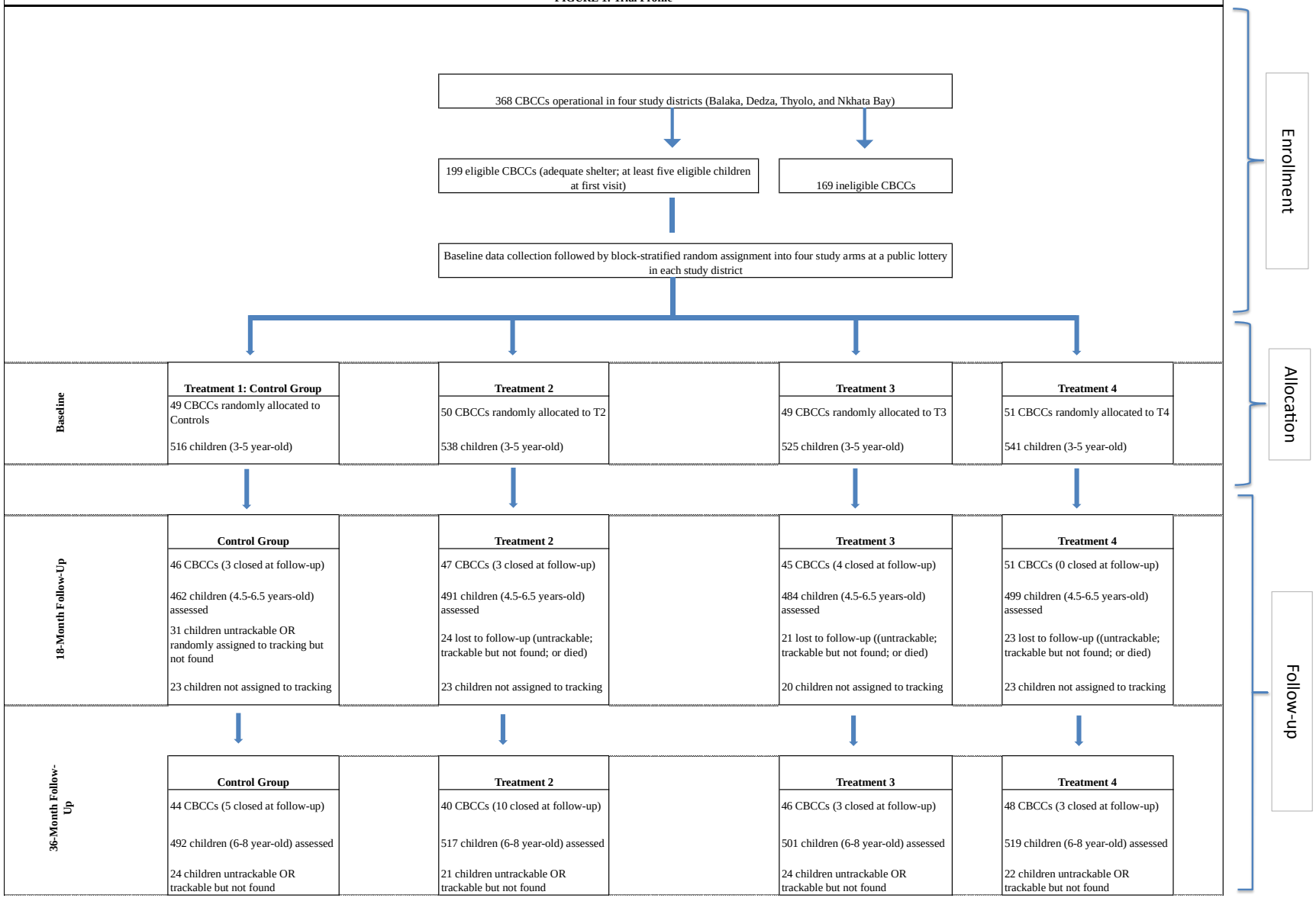

TABLE 1: Baseline Balance (Child-Level Variables)

Variable	Mean (standard deviation) for control group	Difference in Means (compared to the control group)			F-test for Equality of Parameters (p-values)			Number of Observations
		T2 (caregiver training)	T3 (T2 + caregiver incentives)	T4 (T2 + parenting training)	T2 = T3	T2 = T4	T3 = T4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Maintaining Attention and Accuracy During a Test (Leiter Sustained Attention)	24.043 (16.482)	0.660 (1.215)	-0.151 (1.239)	-0.364 (1.319)	0.521	0.446	0.876	2,116
Word Comprehension / Vocabulary (Peabody Picture Vocabulary Test)	24.595 (5.721)	0.609 (0.668)	0.306 (0.500)	0.166 (0.493)	0.667	0.526	0.796	2,116
Language Skills (Malawi Developmental Assessment Tool)	9.497 (3.356)	-0.306 (0.292)	-0.183 (0.267)	-0.369 (0.276)	0.682	0.837	0.513	2,109
Fine Motor / Perception Skills (Malawi Developmental Assessment Tool)	10.232 (3.260)	-0.286 (0.313)	0.001 (0.281)	-0.284 (0.350)	0.370	0.995	0.425	2,116
Male	0.437 (0.496)	0.024 (0.027)	0.026 (0.025)	0.008 (0.023)	0.959	0.525	0.440	2,120
Height-for-Age z-score	-1.667 (1.069)	-0.049 (0.086)	-0.017 (0.087)	0.068 (0.094)	0.685	0.176	0.326	2,110
Age (months)	48.027 (6.684)	-0.193 (0.478)	0.324 (0.478)	0.158 (0.472)	0.294	0.470	0.733	2,120
Chi-Squared Test for Joint Orthogonality of All Variables (p-value)		0.297	0.717	0.539	0.927	0.610	0.903	

Notes: .01 - ***; .05 - **; .1 - *; Cross-sectional OLS regressions at child level using baseline data with standard errors (SEs) between parentheses. SEs are clustered at the CBCC level and observations are weighted using sampling weights. The last row shows the p-values of a series of Joint Orthogonality Tests: we first estimate a Multinomial Logit where the dependent variable is the lottery group, the explanatory variables are the variables in this table, and the base group is the control group. Then, for column (2), we test the null that the coefficients of explanatory variables that refer to T2 are jointly zero, and analogously for columns (3)-(4). For column (5), we reestimate the Multinomial Logit using T2 as a base group, and test the null that the coefficients that refer to T3 are jointly zero, and analogously for columns (6)-(7).

TABLE 2: Attrition (Child Level)

		Dependent Variable: Binary Indicator for Child Lost to Follow-Up			
		18-Month Follow-Up		36-Month Follow-Up	
	Variable	(1)	(2)	(3)	(4)
	T2 (Caregiver training)	-0.016	0.311*	0.004	0.213
		(0.015)	(0.161)	(0.014)	(0.177)
	T3 (T2 + Caregiver incentives)	-0.007	0.266	-0.000	0.090
		(0.017)	(0.213)	(0.013)	(0.166)
	T4 (T2 + Parenting training)	-0.001	0.304	-0.007	0.189
		(0.018)	(0.203)	(0.012)	(0.181)
Child variables	Leiter Sustained Attention		0.015		0.025***
			(0.012)		(0.008)
	Peabody Picture Vocabulary Test		0.018		0.001
			(0.020)		(0.013)
	Malawi Developmental Assessment Tool: Language Skills		-0.021		0.009
			(0.017)		(0.010)
	Malawi Developmental Assessment Tool: Fine Motor / Perception Skills		-0.015		-0.022**
			(0.014)		(0.010)
	Male		-0.005		0.010
			(0.023)		(0.014)
	Age (months)		0.006***		0.002
			(0.002)		(0.003)
	Height-for-Age z-score		-0.015		-0.009
			(0.015)		(0.010)
Interactions: T2 Dummy x Child variables	T2 x Leiter Sustained Attention		-0.001		-0.051**
			(0.017)		(0.025)
	T2 x Peabody Picture Vocabulary Test		-0.016		0.003
			(0.021)		(0.016)
	T2 x Malawi Developmental Assessment Tool: Language Skills		0.030		-0.031
			(0.021)		(0.020)
	T2 x Malawi Developmental Assessment Tool: Fine Motor / Perception Skills		0.020		0.004
			(0.020)		(0.019)
	T2 x Male		-0.015		-0.003
			(0.028)		(0.026)
	T2 x Age (months)		-0.007*		-0.004
			(0.003)		(0.004)
	T2 x Height-for-Age z-score		0.009		0.026
			(0.020)		(0.016)

TABLE 2: Attrition (Child Level) - CONTINUED

	Variable	18-Month Follow-Up (1)	18-Month Follow-Up (2)	36-Month Follow-Up (3)	36-Month Follow-Up (4)
Interactions: T3 Dummy x Child variables	T3 x Leiter Sustained Attention		-0.025		-0.022*
			(0.019)		(0.012)
	T3 x Peabody Picture Vocabulary Test		-0.016		-0.022
			(0.030)		(0.017)
	T3 x Malawi Developmental Assessment Tool: Language Skills		0.024		-0.006
			(0.023)		(0.015)
	T3 x Malawi Developmental Assessment Tool: Fine Motor / Perception Skills		-0.002		0.013
			(0.021)		(0.019)
	T3 x Male		-0.011		0.002
			(0.037)		(0.031)
	T3 x Age (months)		-0.006		-0.002
			(0.004)		(0.003)
	T3 x Height-for-Age z-score		0.007		0.029
			(0.021)		(0.018)
Interactions: T4 Dummy x Child variables	T4 x Leiter Sustained Attention		0.015		-0.029**
			(0.022)		(0.013)
	T4 x Peabody Picture Vocabulary Test		-0.022		0.005
			(0.022)		(0.016)
	T4 x Malawi Developmental Assessment Tool: Language Skills		0.017		0.002
			(0.024)		(0.021)
	T4 x Malawi Developmental Assessment Tool: Fine Motor / Perception Skills		0.010		0.030*
			(0.022)		(0.017)
	T4 x Male		0.007		-0.010
			(0.035)		(0.027)
	T4 x Age (months)		-0.006		-0.004
			(0.004)		(0.004)
	T4 x Height-for-Age z-score		0.016		0.007
			(0.040)		(0.022)
	Mean (standard deviation) of dependent variable for the control group		0.062 (0.242)		0.046 (0.210)
	Joint F-test of Baseline Controls (minus interactions) - p-value		0.314		0.023**
	Joint F-test of Interactions - p-value with T2:		0.363		0.338
	Joint F-test of Interactions - p-value with T3:		0.793		0.180
	Joint F-test of Interactions - p-value with T4:		0.178		0.147
	Number of observations	2,035	2,035	2,120	2,120

Notes: .01 - ***; .05 - **; .1 - *; Cross-sectional OLS regressions using 18- and 36-month attrition data and baseline variables. Regressions are at the child level, standard errors are clustered by CBCC, and sampling weigths are used (at the 18-month follow-up, 42 randomly tracked observations are reweighted and 85 randomly untracked observations are not included). All regressions control for district-bin fixed effects. All test scores and Height-for-Age z-score are standardized by using (weighted) means and standard deviations from the control group at baseline. We replaced the missing values of index variables and Height-for-Age z-score with their (weighted) averages at baseline for the overall sample.

TABLE 3: Impacts on Child Assessments - 18-Month Follow-Up

		Dependent Variable: Malawi Developmental Assessment Tool Score					
		Total		Language Skills		Fine Motor / Perception Skills	
		(1)	(2)	(3)	(4)	(5)	(6)
T2 (caregiver training)		-0.092	-0.063	-0.041	-0.031	-0.134*	-0.107*
		(0.080)	(0.060)	(0.079)	(0.063)	(0.074)	(0.059)
T3 (T2 + caregiver incentives)		0.013	-0.003	0.088	0.085	-0.081	-0.115*
		(0.081)	(0.067)	(0.084)	(0.072)	(0.075)	(0.065)
T4 (T2 + parenting training)		0.115	0.126*	0.183**	0.185**	0.008	0.012
		(0.086)	(0.067)	(0.087)	(0.071)	(0.075)	(0.061)
Lagged Dependent Variable (Baseline)			0.510***		0.426***		0.444***
			(0.033)		(0.031)		(0.034)
F-test for Equality of Parameters (p-value)	T2=T3	0.191	0.367	0.131	0.106	0.454	0.889
	T2=T4	0.007***	0.002***	0.005***	0.001***	0.033**	0.049**
	T3=T4	0.195	0.058*	0.266	0.168	0.199	0.045**
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes	No	Yes
Number of observations		1,936	1,936	1,936	1,936	1,936	1,936

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using standardized test scores at the 18-month follow-up and baseline covariates with standard errors (SEs) in parentheses. SEs are clustered at the CBCC level and observations are weighted using sampling weights and tracking weights (for 42 observations randomly assigned to tracking).

TABLE 4: Impacts on Child's Behavioral Problems - 18-Month Follow-Up

		Dependent variable: Strengths and Difficulties Questionnaire Score			
		Total Difficulties (Inverted)		Prosocial	
Variables		**(1)**	**(2)**	**(3)**	**(4)**
T2 (caregiver training)		-0.039	-0.063	0.098	0.114
		(0.065)	(0.061)	(0.079)	(0.077)
T3 (T2 + caregiver incentives)		0.098	0.088	0.018	0.026
		(0.062)	(0.062)	(0.083)	(0.081)
T4 (T2 + parenting training)		0.105*	0.072	0.261***	0.252***
		(0.056)	(0.053)	(0.080)	(0.078)
Lagged Dependent Variable (Baseline)			0.317***		0.142***
			(0.026)		(0.027)
Missing Lagged Dependent Variable (Baseline)			-0.076		0.032
			(0.099)		(0.086)
F-test for Equality of Parameters (p-value)	T2=T3	0.037**	0.019**	0.325	0.292
	T2=T4	0.019**	0.020**	0.028**	0.056*
	T3=T4	0.897	0.777	0.003***	0.006***
District-bin Fixed Effects?		Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes
Number of observations		1,938	1,938	1,938	1,938

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using Baseline and 18-month data on the Strengths and Difficulties Questionnaire. Standard errors in parentheses are clustered at the CBCC level. Strengths and Difficulties Questionnaire: Total Difficulties Score had its score inverted so that a higher score is better. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up). We replaced the missing values of lagged dependent variable with its (weighted) average for the overall sample at baseline. For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

TABLE 5: Impacts on Parenting Quality - 18-Month Follow-Up

		Dependent variable:							
				Parenting Quality Subcomponents:					
		Parenting Quality Index		Stimulation Index		Positive Practices Index		Parenting Stress Index (Inverted)	
Variables		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
T2 (caregiver training)		-0.023	-0.018	-0.043	-0.022	-0.023	-0.028	0.018	-0.004
		(0.078)	(0.079)	(0.072)	(0.073)	(0.067)	(0.067)	(0.078)	(0.078)
T3 (T2 + caregiver incentives)		0.104	0.128	0.043	0.056	0.046	0.052	0.090	0.097
		(0.081)	(0.082)	(0.085)	(0.085)	(0.067)	(0.063)	(0.076)	(0.077)
T4 (T2 + parenting training)		0.267***	0.258***	0.294***	0.294***	0.104	0.097	0.088	0.078
		(0.074)	(0.073)	(0.075)	(0.073)	(0.064)	(0.061)	(0.072)	(0.073)
Baseline Control Variable (Baseline)			0.247***		0.246***		0.179***		-0.135***
			(0.028)		(0.030)		(0.025)		(0.027)
F-test for Equality of Parameters - p-value	T2=T3	0.109	0.072*	0.287	0.325	0.255	0.177	0.300	0.140
	T2=T4	0.000***	0.000***	0.000***	0.000***	0.024**	0.027**	0.281	0.213
	T3=T4	0.032**	0.088*	0.002***	0.003***	0.327	0.415	0.980	0.757
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes	No	Yes	No	Yes
Number of observations		1,938	1,938	1,937	1,937	1,934	1,934	1,938	1,938

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using baseline and 18-month data on Stimulation, Positive Practices, Parenting Stress Index and The Center for Epidemiologic Studies Depression Scale. Standard errors in parentheses are clustered at the CBCC level. The Parenting Stress Index had its score inverted so that a higher score is better. The variable referred to as 'baseline control variable' is the lagged (baseline) value of the dependent variable, except in column (8), where it is the baseline value of the Center for Epidemiologic Studies Depression Scale. We use all children for whom the dependent variable is available for the 18-month follow-up. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up). For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

TABLE 6: Impacts on CBCC Outcomes - 18-Month Follow-Up

	Dependent variable:				
	Classroom Observation Index		Total Enrollment (reported)		Number of STC-trained Teachers
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
T2 (caregiver training)	0.554**	0.555**	10.989	10.892**	1.492***
	(0.278)	(0.279)	(6.725)	(5.318)	(0.138)
T3 (T2 + caregiver incentives)	1.214***	1.205***	7.667	6.760	1.612***
	(0.281)	(0.286)	(6.815)	(5.390)	(0.140)
T4 (T2 + parenting training)	1.006***	0.996***	17.727***	13.418**	1.554***
	(0.270)	(0.277)	(6.534)	(5.190)	(0.135)
Lagged Dependent Variable (Baseline)		0.014		0.640***	
		(0.076)		(0.071)	
Mean (standard deviation) of dependent variable for the control group	0.000		63.289		0.000
	(1.000)		(25.003)		(0.000)
F-test for Equality of Parameters - p-value T2=T3	0.019**	0.024**	0.623	0.440	0.389
F-test for Equality of Parameters - p-value T2=T4	0.094*	0.112	0.297	0.622	0.645
F-test for Equality of Parameters - p-value T3=T4	0.446	0.445	0.129	0.205	0.668
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable?	No	Yes	No	Yes	No
Number of observations	189	189	187	187	189

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the CBCC level using baseline and 18-month data on Classroom Observations, Enrollment and STC training of teachers. Standard errors in parentheses. Classroom Observation Index is obtained by weighting underlying variables with Inverse Covariance Weights (ICW) as in Casey et al. (2012) and described in detail in the Appendix. For each round, indices are stardardized using the mean and standard deviation for the control group in that round. We use all CBCCs for whom the relevant variables are available for baseline and the 18-Month Follow-Up. STC stands for 'Save the Children'.

TABLE 7: Effect of Parenting and Classroom Quality on Child Outcomes (Instrumental Variables) - 18-Month Follow-Up

PANEL A: Parenting Quality	**Dependent variable:**							
Sample restricted to T2 (caregiver training) and T4 (T2 + parenting training)	**Malawi Developmental Assessment Tool Score**				**Strengths and Difficulties Questionnaire Score**			
	Language Skills		**Fine Motor / Perception Skills**		**Total Difficulties (Inverted)**		**Prosocial**	
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Second Stage								
Parenting Quality Index (Midline)	0.717***	0.711***	0.488**	0.423**	0.479**	0.422**	0.617***	0.540**
	(0.226)	(0.205)	(0.194)	(0.198)	(0.187)	(0.178)	(0.222)	(0.237)
Lagged Dependent Variable (Baseline)		0.336***		0.344***		0.249***		0.064
		(0.050)		(0.068)		(0.046)		(0.046)
Missing Lagged Dependent Variable (Baseline)						0.020		-0.265**
						(0.171)		(0.113)
First Stage	**Dependent variable: Parenting Quality Index**							
T4	0.284***	0.282***	0.284***	0.278***	0.274***	0.279***	0.274***	0.268***
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.059)	(0.058)	(0.058)
Lagged Dependent Variable (Baseline)		0.170***		0.217***		0.149***		0.079*
		(0.039)		(0.048)		(0.037)		(0.042)
Missing Lagged Dependent Variable (Baseline)						0.312		0.288
						(0.261)		(0.266)
F-test for Significance of T4 (F-stat)	23.75	23.61	23.75	22.67	22.10	22.03	22.10	20.99
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	988	988	988	988	996	996	996	996

TABLE 7: Effect of Parenting and Classroom Quality on Child Outcomes (Instrumental Variables) - 18-Month Follow-Up (CONTINUED)

PANEL B: Classroom Quality	**Dependent variable:**							
Sample restricted to Control and T3 (T2 + caregiver incentives)	**Malawi Developmental Assessment Tool Score**				**Strengths and Difficulties Questionnaire Score**			
	Language Skills		**Fine Motor / Perception Skills**		**Total Difficulties (Inverted)**		**Prosocial**	
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Second Stage								
Classroom Observation Index (Midline)	0.063	0.068	-0.054	-0.070	0.101**	0.089*	-0.039	-0.055
	(0.050)	(0.046)	(0.051)	(0.047)	(0.045)	(0.051)	(0.052)	(0.049)
Lagged Dependent Variable (Baseline)		0.354***		0.454***		0.338***		0.168***
		(0.037)		(0.033)		(0.038)		(0.035)
Missing Lagged Dependent Variable (Baseline)						-0.067		-0.091
						(0.271)		(0.140)
First Stage	**Dependent variable: Classroom Quality Index**							
T3	1.307***	1.313***	1.307***	1.313***	1.306***	1.301***	1.306***	1.301***
	(0.162)	(0.156)	(0.162)	(0.156)	(0.163)	(0.157)	(0.163)	(0.157)
Lagged Dependent Variable (Baseline)		0.001		-0.019		0.026		0.017
		(0.021)		(0.026)		(0.019)		(0.017)
Missing Lagged Dependent Variable (Baseline)						-0.183		-0.183
						(0.312)		(0.315)
F-test for Significance of T3 (F-stat)	64.91	70.54	64.91	70.43	64.38	68.53	64.38	68.40
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	876	876	876	876	873	873	873	873

Notes: .01 - ***; .05 - **; .1 - *. OLS regressions at child level using baseline and 18-month data. Standard errors in parentheses are clustered at the CBCC level. Strengths and Difficulties Questionnaire: Total Difficulties Score had its score inverted so that a higher score is better. We use all children for whom the relevant variables are available. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up). For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

TABLE 8: Impacts on Child Assessments - 36-Month Follow-Up

	Dependent variable: Assessment Score							
	Peabody Picture Vocabulary Test		Leiter Sustained Attention		Kaufman Assessment Battery for Children		Early Grade Math Assessment	
Variables	**(1)**	**(2)**	**(1)**	**(2)**	**(5)**	**(6)**	**(7)**	**(8)**
T2 (caregiver training)	0.077	0.057	-0.036	-0.038	0.015	0.034	-0.101	-0.083
	(0.095)	(0.092)	(0.079)	(0.076)	(0.085)	(0.085)	(0.075)	(0.065)
T3 (T2 + caregiver incentives)	0.161	0.146	0.044	0.023	0.051	0.046	-0.028	-0.031
	(0.109)	(0.104)	(0.075)	(0.068)	(0.085)	(0.086)	(0.075)	(0.066)
T4 (T2 + parenting training)	0.113	0.108	-0.005	-0.020	0.046	0.081	-0.040	-0.021
	(0.103)	(0.099)	(0.092)	(0.084)	(0.091)	(0.089)	(0.074)	(0.065)
Lagged Dependent Variable (Baseline)		0.203***		0.363***				
		(0.023)		(0.032)				
Malawi Developmental Assessment Tool: Fine Motor / Perception Skills (Baseline)						0.373***		0.384***
						(0.038)		(0.038)
F-test for Equality of Parameters - p-value T2=T3	0.377	0.333	0.276	0.383	0.648	0.887	0.337	0.443
F-test for Equality of Parameters - p-value T2=T4	0.686	0.556	0.692	0.815	0.690	0.553	0.405	0.366
F-test for Equality of Parameters - p-value T3=T4	0.635	0.697	0.521	0.536	0.956	0.653	0.868	0.883
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Control Variable and Age Dummies?	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	2,009	2,009	2,029	2,029	2,027	2,027	2,027	2,027

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at child level using baseline and 36-month data on scores. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the relevant variables are available. We weight observations using sampling weights. For each round, scores are stardardized using the (weighted) mean and standard deviation for the control group in that round. For Kaufman Assessment Battery for Children and Early Grade Math Assessment, total scores were obtained by constructing a weighted average of the subscale scores of each assessment (for Kaufman Assessment Battery for Children: Hand Movements, Triangles, and Number Recall / for Early Grade Math Assessment: Number Recognition, Quantity Discrimination, and Addition), where the weights were obtained using Inverse Covariance Weighting, as in Casey et al. (2012) and described in detail in the Appendix.

TABLE 9: Impacts on Child's Behavioral Problems - 36-Month Follow-Up

		Dependent variable: Strengths and Difficulties Questionnaire Score			
		Total Difficulties (Inverted)		Prosocial	
Variables		**(1)**	**(2)**	**(3)**	**(4)**
T2 (caregiver training)		0.124**	0.099*	-0.059	-0.052
		(0.060)	(0.058)	(0.068)	(0.067)
T3 (T2 + caregiver incentives)		0.081	0.100	-0.005	-0.012
		(0.071)	(0.069)	(0.072)	(0.070)
T4 (T2 + parenting training)		-0.046	-0.048	0.020	0.002
		(0.063)	(0.060)	(0.074)	(0.071)
Lagged Dependent Variable (Baseline)			0.244***		0.107***
			(0.025)		(0.023)
Missing Lagged Dependent Variable (Baseline)			0.112		-0.209*
			(0.089)		(0.107)
F-test for Equality of Parameters (p-value)	T2=T3	0.511	0.985	0.386	0.525
	T2=T4	0.003***	0.006***	0.214	0.383
	T3=T4	0.057*	0.025**	0.720	0.833
District-bin Fixed Effects?		Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes
Number of observations		2,022	2,022	2,022	2,022

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at children level using baseline and 36-month data on the Strengths and Difficulties Questionnaire. Standard errors in parentheses are clustered at the CBCC level. Strengths and Difficulties Questionnaire: Total Difficulties Score had its score inverted so that a higher score is better. We weight observations using sampling weights. For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

TABLE 10: Impacts on Parenting Quality - 36-Month Follow-Up

		Dependent variable:			
		Stimulation Index		Positive Practices Index	
Variables		**(1)**	**(2)**	**(3)**	**(4)**
T2 (caregiver training)		-0.045	-0.036	-0.051	-0.064
		(0.076)	(0.073)	(0.071)	(0.072)
T3 (T2 + caregiver incentives)		0.062	0.061	0.002	-0.013
		(0.082)	(0.081)	(0.067)	(0.068)
T4 (T2 + parenting training)		0.172**	0.164**	-0.067	-0.071
		(0.082)	(0.077)	(0.067)	(0.067)
Lagged Dependent Variable (Baseline)			0.207***		0.110***
			(0.030)		(0.032)
F-test for Equality of Parameters - p-value	T2=T3	0.138	0.180	0.454	0.444
	T2=T4	0.003***	0.005***	0.821	0.917
	T3=T4	0.150	0.186	0.315	0.379
District-bin Fixed Effects?		Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes
Number of observations		2,033	2,033	2,030	2,030

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using baseline and 36-month data on Stimulation and Positive Practices. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the dependent variable is available for the 36-Month Follow-Up. Even columns control for the baseline value of the dependent variable. We weight observations using sampling weights. For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

TABLE 11: Impacts on CBCC Outcomes - 36-Month Follow-Up

	Dependent variable:				
	Classroom Observation Index		Total Enrollment (reported)		Number of STC-trained Teachers
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
T2 (caregiver training)	0.284	0.301	4.597	5.845	1.272***
	(0.303)	(0.305)	(7.963)	(6.795)	(0.143)
T3 (T2 + caregiver incentives)	0.335	0.324	-1.589	1.049	1.233***
	(0.290)	(0.292)	(7.624)	(6.514)	(0.137)
T4 (T2 + parenting training)	0.167	0.135	10.259	7.426	1.237***
	(0.291)	(0.298)	(7.645)	(6.534)	(0.137)
Lagged Dependent Variable (Baseline)		0.046		0.614***	
		(0.089)		(0.089)	
Mean (standard deviation) of dependent variable for the control group	0.000		67.545		0.000
	(1.000)		(42.738)		(0.000)
F-test for Equality of Parameters - p-value T2=T3	0.864	0.940	0.432	0.475	0.782
F-test for Equality of Parameters - p-value T2=T4	0.697	0.601	0.476	0.816	0.801
F-test for Equality of Parameters - p-value T3=T4	0.557	0.517	0.118	0.326	0.981
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable?	No	Yes	No	Yes	No
Number of observations	178	178	178	178	178

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the CBCC level using baseline and 36-month data on Classroom Observations, Enrollment and STC training of teachers. Standard errors in parentheses. Classroom Observation Index is obtained by weighting underlying variables with Inverse Covariance Weights (ICW) as in Casey et al. (2012) and described in detail in the Appendix. For each round, indices are stardardized using the mean and standard deviation for the control group in that round. We use all CBCCs for whom the relevant variables are available for baseline and the 36-Month Follow-Up. STC stands for 'Save the Children'.

TABLE 12: Heterogeneity of Impacts - 18-Month Follow-Up

		Dependent variable:					
		Malawi Developmental Assessment Tool Score			Strengths and Difficulties Questionnaire Score		Parenting Quality Index
		Total	Language Skills	Fine Motor / Perception	Total Difficulties	Prosocial	
Variables		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
T2 (caregiver training)		-0.0557	-0.025	-0.096	-0.0557	0.115	-0.0195
		(0.0606)	(0.063)	(0.061)	(0.0615)	(0.075)	(0.0770)
T3 (T2 + caregiver incentives)		-0.00226	0.078	-0.103	0.102*	0.015	0.120
		(0.0670)	(0.071)	(0.066)	(0.0601)	(0.079)	(0.0790)
T4 (T2 + parenting training)		0.112	0.172**	-0.000	0.0754	0.259***	0.244***
		(0.0688)	(0.071)	(0.063)	(0.0516)	(0.073)	(0.0713)
Lagged Dependent Variable (Baseline)		0.447***	0.350***	0.420***	0.302***	0.206***	0.257***
		(0.0514)	(0.058)	(0.049)	(0.0424)	(0.066)	(0.0443)
Age in Months (Baseline)		0.0239***	0.030***	0.024**	0.00633	0.010	-0.00785
		(0.00868)	(0.008)	(0.009)	(0.00673)	(0.006)	(0.00837)
Primary Caregiver has a Primary School Leaving Certificate (Baseline)		0.214**	0.167	0.249**	-0.0173	-0.000	0.125
		(0.101)	(0.112)	(0.102)	(0.112)	(0.102)	(0.119)
Height-for-Age z-score (Baseline)		0.148***	0.172***	0.122**	0.0270	0.102**	0.0963**
		(0.0479)	(0.045)	(0.053)	(0.0517)	(0.049)	(0.0395)
T2 x Primary Caregiver has a Primary School Leaving Certificate (Baseline)		-0.0770	-0.070	-0.009	0.346**	0.088	0.185
		(0.142)	(0.144)	(0.163)	(0.172)	(0.201)	(0.171)
T3 x Primary Caregiver has a Primary School Leaving Certificate (Baseline)		-0.0482	-0.144	0.074	0.281**	0.009	0.119
		(0.137)	(0.151)	(0.145)	(0.142)	(0.162)	(0.145)
T4 x Primary Caregiver has a Primary School Leaving Certificate (Baseline)		0.268	0.510*	-0.044	0.0806	0.198	0.160
		(0.194)	(0.267)	(0.155)	(0.140)	(0.166)	(0.146)
T2 x Height-for-Age z-score (Baseline)		-0.127*	-0.103	-0.124	-0.0586	-0.027	-0.109
		(0.0668)	(0.063)	(0.078)	(0.0706)	(0.064)	(0.0672)
T3 x Height-for-Age z-score (Baseline)		-0.0836	-0.087	-0.066	-0.0406	-0.085	-0.161***
		(0.0566)	(0.056)	(0.068)	(0.0659)	(0.072)	(0.0577)
T4 x Height-for-Age z-score (Baseline)		-0.0828	-0.122**	-0.028	0.0655	-0.144**	0.0219
		(0.0599)	(0.057)	(0.070)	(0.0909)	(0.071)	(0.0895)
F-test for Equality of Parameters - p-value	T2=T3	0.413	0.149	0.915	0.012**	0.218	0.074*
	T2=T4	0.007***	0.002***	0.109	0.019**	0.045**	0.000***
	T3=T4	0.094*	0.192	0.110	0.610	0.003***	0.090*
Joint F-test of Interactions of Controls with (T2, T3, T4) - p-value	T2	0.201	0.402	0.373	0.155	0.392	0.075*
	T3	0.491	0.443	0.557	0.320	0.701	0.021**
	T4	0.287	0.055*	0.961	0.943	0.099*	0.678
Number of observations		1,928	1,928	1,928	1,930	1,930	1,930

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at children level using baseline and 18-Month data. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the dependent variable is available for the 18-month follow-up. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up). Strengths and Difficulties Questionnaire: Total Difficulties Score had its score inverted so that a higher score is better. For brevity, the coefficient estimates for interactions of treatment dummies with the Lagged Dependent Variable and Age in Months are ommited from the table (but included in the regression analysis). For each round, scores and indices are standardized using the (weighted) mean and standard deviation for the control group in that round. District-bin fixed effects are included.

Appendix: Child Assessments, Primary Caregiver Surveys, and CBCC Observation Tools

Child measures

The battery included the following measures:

1. *Malawi Developmental Assessment Tool (MDAT)(Gladstone et al., 2010)* a test created and validated specifically for use in rural Malawi with Chichewa-speaking children 0-7 years of age.[25] The MDAT includes four subscales to assess *Language, Fine Motor/Perception, Gross Motor* and *Personal-Social* skills. The majority of the items were designed to be administered directly to the child using locally available materials. For example, a *Fine Motor/Perception* question asks a child to copy a pattern using bottle tops (e.g., a square pattern with alternating green and red bottle tops), which are commonly available and used for various games. A *Language* question asking a child to explain what objects are used for includes showing the child familiar objects widely found in rural households, such as a small, homemade broom (used for sweeping), and a matchbox (containing matches, used for lighting stove). For our study, only the *Language* and *Fine Motor/Perception* subscales were administered as these assessed skills most closely related to the interventions. We supplemented the original MDAT items in these subscales with questions that reflected the content of the preschool teacher and caregiver education trainings, including items asking about copying letters and naming colors. Items were scored as pass or fail, and a total summed score was calculated overall, and for each subscale. A *Total* score was also computed by adding the *Language* and *Fine Motor/Perception* subscales. The *Language*, *Fine Motor/Perception*, and *Total* scores were then standardized at each round (Baseline and 18-Month Follow-Up) using the control group's mean and standard deviation.
2. *Peabody Picture Vocabulary Test - IV (PPVT-IV, (Dunn, 1965)* a test of receptive vocabulary that measures comprehension of words through picture identification for use with people 2.5 years

[25] The original MDAT materials can be accessed online at http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000273.

and older.[26] The PPVT has been widely used throughout the world for assessing the effects of various interventions on child language, including Mozambique (Spanish version of PPVT; (Martinez et al., 2012)) and Madagascar (Fernald et al., 2009). In administering the PPVT, the child was shown a page with four pictures, and then asked to point to or touch the picture (stimulus word) named by the enumerator. Specific items (both words and pictures) were modified for use in Malawi. For example, we replaced "apple" with "papaya," a fruit that is well known throughout the country, and was estimated to be of similar difficulty as the word "apple" would be in the United States. Another example was changing a stimulus word from "tornado" to "whirlwind," which was more familiar to children for describing the accompanying picture. All items administered were translated and back translated (English-Chichewa-English). The test publishers approved changes made to the items. Items were scored as pass or fail, and a summed score was calculated. This score was then standardized at each round (Baseline and 36-Month Follow-Up) using the control group's mean and standard deviation.

3. *The Leiter-R Sustained Attention (LSA) task (Roid & Miller, 1997),* a language-free measure that assesses how well children can continue to maintain attention and accuracy during a timed visual search task.[27] To administer this test, the child was taught to recognize a target picture (e.g., butterfly), and then asked to mark all of the exact same targets (butterflies) in an array of many pictures, in a fixed amount of time. There are three sets of age-specific tasks (2-3 years old; 4-5 years old; 6+ years old), and each set included four different tasks of increasing difficulty. The measure has successfully detected group differences in performance in Madagascar (Fernald et al., 2011). Total adjusted scores were determined by subtracting the numbers of errors from the number of correct responses. The scores were then standardized at each round (Baseline and 36-

[26] The PPVT can be purchased through http://www.pearsonclinical.com/language/products/100000501/peabody-picture-vocabulary-test-fourth-edition-ppvt-4.html.

[27] The Leiter-R test materials can be purchased through http://www.stoeltingco.com/psychologicaltesting/intellectual-cognititve/nonverbal/leiter-3-kit-in-rolling-backpack.html.

Month Follow-Up) using the control group's mean and standard deviation.

4. *The Strengths and Difficulties Questionnaire (SDQ) (Goodman, 2001; Woerner et al., 2004),* a 25-item, parent-report questionnaire that screens for both behavioral problems and pro-social (positive) behaviors. The SDQ includes four problem-behavior subscales (*Emotional Symptoms, Conduct Problems, Hyperactivity/Inattentiveness,* and *Peer Relationship Problems*), along with a pro-social subscale. Examples of problematic behaviors include "Child is easily distracted, concentration wanders," and "Child is often picked on or bullied." Pro-social behavior items include "Child is kind to younger children," and "Child shares readily with other children." All items were translated, back translated, and approved by the test author.[28] The SDQ has been used in several African countries, including Kenya (Oburu, 2005) and South Africa (Cluver et al., 2007). For our administration, all items were read to the caregiver, who provided a response indicating her degree of agreement with the statement. Scores were determined for the four behavior problem subscales, which were aggregated into a total difficulties (problem) score, and the pro-social subscale separately. The total difficulties score and the pro-social scores were then standardized at each round (baseline, 18-, and 36-month follow-up) using the control group's mean and standard deviation. Finally, the total difficulties index was multiplied by -1 so that a higher score indicates fewer difficulties.
5. *Kaufman Assessment Battery-Children, 2nd Edition (KABC-II) (Kaufman & Kaufman, 2004),* a suite of cognitive tests from which we adopted three tasks that performed well in piloting: *Hand Movements*, *Number Recall*, and *Triangles*.[29] *Hand Movements* is a non-verbal, short-term motor memory task requiring children to copy increasingly difficult hand movement sequences. For this test, the assessor demonstrated the movements by touching the table or floor as prescribed (e.g.

[28] The Chichewa version of the SDQ can be found online here: http://www.sdqinfo.com/py/sdqinfo/b3.py?language=Chichewa. The website also provides instructions for scoring.

[29] The Kaufman scales have been used in Kenya (Holding et al., 1999), Senegal (Boivin, 2002), and Uganda (Bangirana et al., 2009). The KABC-II test materials can be purchased through http://www.pearsonclinical.com/psychology/products/100000088/kaufman-assessment-battery-for-children-second-edition-kabc-ii.html.

fist, palm, fist), and then the child tried to replicate the sequence in the proper order. *Number Recall* is a short-term auditory memory task requiring children to repeat a series of increasingly difficult number sequences (e.g., 2-4, 3-9-5, 4-1-9-2, etc.) spoken by the assessor. As children learn numbers in English, no translation was required. *Triangles* is a non-verbal problem-solving task that requires children to complete increasingly complex patterns and figures with plastic and foam triangle shapes. For this task administration, the enumerator either modeled how to make the object (e.g., making a "car" using square and round plastic pieces), or showed the child a picture to copy (per test instructions). For each of the three tasks, the test ended when the child failed three consecutive items. A total score of passed items was calculated for each task.

In order to aggregate the scores of *Hand Movements*, *Number Recall*, and *Triangles* into one single Kaufman score, we used the Inverse Covariance Weighting (ICW) methodology in Casey et al. (2012), which consists of the following steps: (*i*) standardizing each of the three scores using the control group's mean and standard deviation; (*ii*) computing the variance-covariance matrix of the standardized variables; (*iii*) generating a weighted average of the standardized variables where the weights are proportional to the sums of the rows of the inverted variance-covariance matrix; (*iv*) standardizing this weighted average using the control group's mean and standard deviation.

6. *Early Grade Mathematics Assessment (EGMA) (Brombacher, 2011)*, a tool developed by the United States Agency for International Development (USAID) to measure early knowledge of numbers and basic math skills, validated in Malawi.[30] A great advantage of the EGMA is that there are Malawian norms available, as well as norms from nearby countries (Kenya, Tanzania, Zambia), allowing for easy comparison and interpretation of scores. Three subscales were administered. The *Number Recognition* task required the child to name 20 one-, two- and three-digit numbers in 60 seconds. The *Quantity Discrimination* subscale had 10 items that asked

[30] The Chichewa EGMA scales can be downloaded at: https://www.eddataglobal.org/countries/index.cfm?fuseaction=pubDetail&ID=29.

children to identify (point to or name) the larger of two numbers; the test ended when the child failed four consecutive items. The *Addition* task included 20 equations, and the child was instructed to say the sum for each item in the maximum time allowed (60 seconds). Passed items for each subscale were summed to create subscale scores.

In order to aggregate the scores of *Number Recognition*, *Quantity Discrimination*, and *Addition* into one single EGMA score, we used the Inverse Covariance Weighting (ICW) methodology in Casey et al. (2012), which is described in detail under *Kaufman Assessment Battery-Children, 2nd Edition (KABC-II)* above.

All enumerators were trained for a minimum of two weeks at each data collection time point, and all followed standardized procedures for administering each measure. Inter-rater reliability, as indicated by the correlation between scores obtained by two different testers for the same child, was estimated by having the enumerators observe and score videotaped administrations. Average inter-rater reliabilities were 0.95 (for MDAT Fine Motor at baseline and Round 2), 0.88 (for MDAT Language at baseline and Round 2), 0.94 (PPVT, baseline), and 0.96 (Triangles, Round 3).

Primary caregiver measures

The following scales were administered to each caregiver:

1. *The Center for Epidemiological Studies, Depression (CESD) (Radloff, 1977)*, a 20-item, self-report scale that assesses the frequency of common depressive symptoms experienced in the past week. The CESD has been widely used with adults throughout the world, including in low- and middle-income countries (e.g. (Baker-Henningham et al., 2005; Black, 2007). Items were translated and back translated. Items included, "Did you feel sad?" "Did you feel that everything you did was an effort?" and "Were you happy?" For administration, all items were read to the caregiver. Higher scores indicated the presence of more depressive symptoms.[31] The final index was obtained by standardizing the scale using the control group's mean and standard deviation.

[31] The CESD can be downloaded through http://cesd-r.com/.

2. *The Parenting Stress Index-Short Form (PSI/SF) (Abidin, 1990),* an adapted 43-item scale that asks caregivers to report on their perceptions of parenting the target child in the study. The PSI/SF consists of three 12-item subscales: *Parental Distress, Parent-Child Dysfunctional Interaction*, and *Difficult Child*. The PSI/SF has been used in South Africa (Allen et al., 2014) and Kenya (Oburu, 2005). Sample items include, "You expected to have closer and warmer feelings for [CHILD] than you do, and this bothers you," "[CHILD] makes more demands on you than most children," and "You often have the feeling you cannot handle things very well." All items were translated and back translated, and changes were made per recommendations by the PSI/SF author. During administration, all items were read to the caregiver, who indicated the degree to which they agreed with the statement. The three subscales were summed to create a Total Stress score. Higher scores indicated more stress related to parenting this child. We added 7 items to also capture positive feelings related to caring for the child, which were interspersed into the PSI/SF, which were not used in calculating the Total Stress score.[32] The final index was obtained by standardizing the Total Stress score using the control group's mean and standard deviation at each round (baseline and 18-month follow-up) and then multiplying it by -1 so that a higher score indicates less parental stress.
3. *Support for Learning and Positive Parenting* (UNICEF, 2010) module was adapted from the UNICEF Multi-Indicator Cluster Surveys (MICS) and other sources (Hamadani et al., 2010; Kariger et al., 2012). Support for learning was determined by both the availability of materials (books, toys etc.) that promote development, as well as activities adults do with children to encourage learning. We expanded the activities to include caregiver-child interactions that were related to the parenting intervention, and which would encourage school readiness. These included helping the child learn numbers and letters; teaching the child the name and use of

[32] The PSI/SF can be purchased through: http://www4.parinc.com/Products/Product.aspx?ProductID=PSI-4:SF.

objects; and helping child with any homework. Typical behavior control strategies or disciplinary techniques were also measured, adapted from the MICS.[33]

Two scores were derived from this module: (*i*) the *Stimulation Index* sums over a set of indicators describing activities that the primary caregiver does with the child (13 variables at baseline and the 18-month follow-up, and 18 variables at the 36-month follow-up); (*ii*) the *Positive Practices Index* sums over a set of six indicators for the use of positive methods to address child behavioral problems. Both the *Stimulation Index* and the *Positive Practices Index* were standardized at each round (baseline, 18-, and 36-month follow-up) using the control group's mean and standard deviation.

We used Inverse Covariance Weighting (ICW) as described in Casey et al. (2012) to generate a *Parenting Quality Index* by aggregating over parenting-related variables (*Stimulation Index*, *Positive Practices Index*, and *Parental Stress Index*) at baseline and the 18-month follow-up. The ICW methodology is described in detail under *Kaufman Assessment Battery-Children, 2nd Edition (KABC-II)* above. The *Parental Quality Index* is not computed at 36-Month Follow-Up since we did not administer the *Parental Stress Index* during that round.

Community-based child care center (CBCC) quality measures

We used two strategies for gathering data on the quality of the CBCCs to estimate impacts of the interventions on the functioning of the centers. These included a questionnaire, completed by the director and teachers; and an observational tool, completed by a pair of enumerators during typical operation of the center. The CBCC questionnaire and observation measures were adapted from the *La Escala de Evaluación de la Calidad Educativa de Centros de Educación Preescolares* (ECCP) (Martínez et al., 2004) from Mexico, and a preschool quality tool used in Cambodia (Rao et al., 2012). Items on both tools were based on theoretical and empirical data recommending assessment of both the *structural* (e.g., building characteristics, water and sanitation facilities, availability of appropriate and diverse learning materials, provision of snacks, safety of environment, schedule of activities) and *process* (e.g., warmth

[33] The UNICEF MICS questionnaires can be accessed at http://mics.unicef.org/tools.

and responsiveness of teachers toward the child, use of teaching strategies that encourage child participation, capacity for group and one-on-one interaction between teachers and children, supervision and disciplinary techniques) qualities of early learning centers.

Items were translated, back translated, reviewed, and adapted by Malawian early child development experts. Some items were also drawn from previous evaluations of CBCCs (Chiuye & Chimombo, 2011; Munthali et al., 2014). Our decision to use these measures, as opposed to other standardized Western tools such as the Early Childhood Environment Rating Scale (ECERS; (Clifford et al., 2010; Harms & Clifford, 1982)) and the Classroom Assessment Scoring System (CLASS; (Pianta et al., 2007)) was based on several factors. The CBCCs are unlike preschools and kindergartens for which the ECERS and CLASS were developed in that they are initiated, run and staffed by community volunteers with minimal education, training and oversight. Thus, the centers are not universal in form or function: they may operate erratically based on the availability of structures, teachers and materials; they tend to serve a wide range of ages, and the emphasis of preparing children for primary school may be secondary to other goals, such as providing a safe place for children to play, and get a meal. Our measures, informed by those developed in low- and middle-income countries, with added information from local experts, strove to measure quality in a contextually relevant manner. While these measures may not be externally valid (e.g., for comparing results across preschool interventions), they are useful for providing tools that can be used within Malawi for evaluating quality of centers. Our items might also be relevant to efforts at creating a universal measure of early learning environment quality.

We used the data collected under the observational tool to develop a *Classroom Observation Index* with the purpose of measuring classroom quality. At each round, the index computes a weighted average of a set of underlying observational variables (31, 34, and 32 variables at baseline, 18-, and 36-month follow-up, respectively) using the Inverse Covariance Weighting (ICW) methodology in Casey et al. (2012), which is described in detail under *Kaufman Assessment Battery-Children, 2nd Edition (KABC-II)* above.

We also used ICW to develop two *Classroom Observation subscales*. For each round, we partition the underlying variables into two groups according to their pre-specified domains. The first group includes variables related to *Structure, Engagement, and Supervision*, while the second group includes variables related to *Literacy and Numeracy, Problem Solving, and Fine and Gross Motor Activities*. We then use ICW to aggregate over the variables within each group, to generate two subscale scores that were used in Appendix Table 8.

Finally, we used *Principal Component Analysis* (PCA) to identify the latent orthogonal factors underlying the classroom observation index. We run PCA separately by round using all available classroom observation variables at each round, and then extract the first two principal components. These principal components are then standardized per round (baseline, 18-, and 36-month follow-up) using the control group's mean and standard deviation. Impacts on these principal components are also presented in Appendix Table 8.

References

Abidin, R. R. (1990). Parenting Stress Index/Short Form. Psychological Assessment Resources: Lutz, FL.

Allen, A. B., et al. (2014). The role of parenting in affecting the behavior and adaptive functioning of young children of HIV-infected mothers in South Africa. *AIDS Behav,* 18(3), 605-616.

Baker-Henningham, H., Powell, C., Walker, S., & Grantham-McGregor, S. (2005). The effect of early stimulation on maternal depression: a cluster randomised controlled trial. *Arch Dis Child,* 90(12), 1230-1234.

Bangirana, P., et al. (2009). A preliminary examination of the construct validity of the KABC-II in Ugandan children with a history of cerebral malaria. . *African Health Sciences,* 9, 188-192.

Black, M. M., Baqui, A. H., Zaman K., McNary, S. W., Le, K., Arifeeen, S. E., Hamadani, J. D., Parveen, M., Yunus, M., & Black, R. E. (2007). Depressive symptoms among rural Bangladeshi mothers: implications for infant development. *Journal of Child Psychology and Psychiatry,*, 48, 764-772.

Boivin, M. J. (2002). Effects of early cerebral malaria on cognitive ability in Senegalese children. *J Dev Behav Pediatr,* 23, 353-364.

Brombacher, A. (2011). Malawi Early Grade Mathematics Assessment (EGMA): National Baseline Report 2010. Malawi: USAID/Malawi and the Ministry of Education, Science and Technology.

Chiuye, G., & Chimombo, J. P. G. (2011). Evaluation of Community Based Quality Early Childhood Development (ECD) ELMA Project in Lilongwe, Dedza and Zomba Districts: Center for Educational Research and Training, University of Malawi.

Clifford, R. M., Reszka, S. S., & Rossbach, H. G. (2010). Reliability and validity of the early childhood environment rating scale. Retrieved September,30, 2013.

Cluver, L., Gardner, F., & Operario, D. (2007). Psychological distress amongst AIDS orphaned children in urban South Africa. *Journal of Child Psychology and Psychiatry,*, 48, 755-763.

Dunn, L. N. (1965). Peabody Picture Vocabulary Test. Nashville, Tennessee: American Guidance Service.

Fernald, L. C., Weber, A., Galasso, E., & Ratsifandrihamanana, L. (2011). Socioeconomic gradients and child development in a very low income population: evidence from Madagascar. *Dev Sci*, 14(4), 832-847.

Fernald, L. C. H., Kariger, P., Engle, P., & Raikes, A. (2009). Examining early child development in low-income countries: a toolkit for the assessment of children in the first five years of life. Washington, DC: The World Bank.

Gladstone, M., et al. (2010). The Malawi Developmental Assessment Tool (MDAT): the creation, validation, and reliability of a tool to assess child development in rural African settings. . *PLoS Med*, 7(5).

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. . *Journal of the American Academy of Child & Adolescent Psychiatry*, 40(11), 1337-1345.

Hamadani, J. D., et al. (2010). Use of family care indicators and their relationship with child development in Bangladesh. *J Health Popul Nutr*, 28(1), 23-33.

Harms, T., & Clifford, R. M. (1982). Assessing preschool environments with the early childhood environment rating scale. *Studies in Educational Evaluation*, 8(3), 261-269.

Holding, P. A., Stevenson, J., Peshu, N., & Marsh, K. (1999). Cognitive sequelae of severe malaria with impaired consciousness. *Trans R Soc Trop Med Hyg*, 93, 529-534.

Kariger, P., et al. (2012). Indicators of family care for development for use in multicountry surveys. *J Health Popul Nutr*, 30(4), 472-486.

Kaufman, A. S., & Kaufman, N. L. (2004). Kaufman Assessment Battery for Children, Second Edition Examiners Manual.

Martínez, J. F., Myers, R., & Linares, M. (2004). ¿Todos los pollos son amarillos?: En búsqueda de la calidad educativa en centros preescolares. [Are all chickens yellow? An investigation into the quality of preschool centers.]. Mexico: Dirección General de Investigación Educativa. Secretaría de Educación Pública.

Martinez, S., Naudeau, S., & Pereira, V. (2012). The promise of preschool in Africa: A randomized impact evaluation of early childhood development in rural Mozambique. Available at http://www.3ieimpact.org/media/filer/2013/04/11/3ie_mozambique_ie001.pdf (Accessed August 9, 2013)

Munthali, A. C., Mvula, P. M., & Silo, L. (2014). Early childhood development: the role of community based childcare centres in Malawi. *SpringerPlus*, 3, 305.

Oburu, P. O. (2005). Caregiving stress and adjustment problems of Kenyan orphans raised by grandmothers. *Infant and Child Development*, 14, 199-210.

Pianta, R., K. , LaParo, K., & Hamre, B. C. (2007). Classroom Assessment Scoring System-CLASS. Baltimore: Brookes.

Radloff, L. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement*, 1, 385-401.

Rao, N., et al. (2012). Is something better than nothing? An evaluation of early childhood programs in Cambodia. *Child Dev*, 83(3), 864-876.

Roid, G. H., & Miller, L. J. (1997). Leiter International Performance Scale - Revised (Leiter-R). Wood Dale: Stoetling

UNICEF. (2010). 2010 UNICEF Multiple Indicator Cluster Survey.

Woerner, W., et al. (2004). The Strengths and Difficulties Questionnaire overseas: evaluations and applications of the SDQ beyond Europe. *Eur Child Adolesc Psychiatry*, 13 Suppl 2, 47-54.

Appendix Table 1: Schedule of Child Assessments

Measure	*Time of Data Collection*		
	Baseline	18-Months	36-Months
Anthropometric measurements (Height & Weight)	X		
Malawi Developmental Assessment Tool (MDAT)	X	X	
Peabody Picture Vocabulary Test (PPVT-IV)	X		X
Leiter-R Sustained Attention (LSA)	X		X
Strengths and Difficulties Questionnaire (SDQ)	X	X	X
KABC-II			X
Early Grade Math Assessment (EGMA)			X

APPENDIX TABLE 2: Baseline Balance (Primary Caregiver-Level Variables)

Variable	Mean (standard deviation) for control group	Difference in Means (compared to the control group)			F-test for Equality of Parameters (p-values)			Number of Observations
		T2 (caregiver training)	T3 (T2 + caregiver incentives)	T4 (T2 + parenting training)	T2 = T3	T2 = T4	T3 = T4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Activities with Children (Stimulation Index)	4.188 (3.035)	-0.335 (0.232)	-0.082 (0.266)	-0.109 (0.259)	0.233	0.267	0.910	2,113
Use of Positive Disciplinary Techniques (Positive Practices Index)	3.174 (1.215)	0.181* (0.094)	0.086 (0.106)	0.146 (0.103)	0.337	0.710	0.578	2,113
Primary Caregiver has a Primary School Leaving Certificate	0.218 (0.413)	-0.011 (0.037)	-0.012 (0.034)	-0.001 (0.033)	0.997	0.774	0.755	2,113
Child's Behavioral Problems (Strengths and Difficulties Questionnaire: Total Difficulties Score)	14.027 (6.241)	-0.666 (0.496)	0.129 (0.484)	-0.190 (0.584)	0.137	0.448	0.604	1,815
Child's Positive Behaviors (Strengths and Difficulties Questionnaire: Prosocial Score)	6.200 (2.368)	-0.480* (0.276)	-0.175 (0.241)	0.002 (0.266)	0.256	0.099*	0.493	1,815
Missing Strengths and Difficulties Questionnaire	0.193 (0.395)	0.002 (0.089)	-0.093 (0.077)	0.006 (0.114)	0.203	0.976	0.339	2,113
Primary Caregiver's Depressive Symptoms (The Center for Epidemiologic Studies Depression Scale)	16.266 (7.220)	-1.150* (0.608)	0.012 (0.552)	-0.492 (0.713)	0.046**	0.371	0.464	2,113
Household Wealth Index	-0.036 (0.916)	-0.086 (0.106)	0.030 (0.106)	0.166 (0.101)	0.282	0.016**	0.191	2,110
Chi-Squared Test for Joint Orthogonality of All Variables (p-value)		0.168	0.990	0.876	0.262	0.167	0.899	

Notes: .01 - ***; .05 - **; .1 - *. Cross-sectional OLS regressions at child level using baseline data with standard errors (SEs) between parentheses. SEs are clustered at the CBCC level and observations are weighted using sampling weights. Following Filmer and Scott (2012), the Household Wealth Index is built using a two-parameter logistic (2PL) Item Response Theory (IRT) model on a set of indicator variables for household assets, infrastructure, and land holdings. The indicator variables include: non-grass roof (e.g. clay, concrete, iron), non-natural floor (e.g. brick, wood), personal source of water (e.g. piped, well), access to electricity, access to a flush toilet, above-median land acreage, and a set of 19 household items such as radio, car and motorcycle. The last row shows the p-values of a series of Joint Orthogonality Tests: we first estimate a Multinomial Logit where the dependent variable is the lottery group, the explanatory variables are the variables in this table (except 'Missing Strengths and Difficulties Questionnaire'), the base group is the control group, and the sample includes all observations for which none of the variables are missing. Then, for column (2), we test the null that the coefficients of explanatory variables that refer to T2 are jointly zero, and analogously for columns (3)-(4). For column (5), we reestimate the Multinomial Logit using T2 as a base group, and test the null that the coefficients that refer to T3 are jointly zero, and analogously for columns (6)-(7).

APPENDIX TABLE 3: Baseline Balance (CBCC-Level Variables)

Variable	Mean (standard deviation) for control group	Difference in Means (compared to the control group)			F-test for Equality of Parameters (p-values)			Number of Observations
		T2 (caregiver training)	T3 (T2 + caregiver incentives)	T4 (T2 + parenting training)	T2 = T3	T2 = T4	T3 = T4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Classrooom Quality (Classroom Observation Index)	0.000 (1.000)	-0.132 (0.315)	0.403 (0.317)	0.706** (0.314)	0.091*	0.008***	0.335	199
Age of the Community-Based Childcare Center (CBCC)	6.878 (5.967)	-0.558 (0.910)	-1.128 (0.919)	-1.172 (0.906)	0.534	0.496	0.961	198
Total Number of 3-4 Year-Old Children in the CBCC (observed)	17.347 (7.970)	-0.687 (2.157)	-0.898 (2.168)	3.712* (2.147)	0.922	0.041**	0.033**	199
Total Enrollment (reported)	57.184 (27.390)	-2.304 (6.081)	-2.184 (6.111)	4.895 (6.051)	0.984	0.233	0.244	199
Average Daily Attendance (reported)	37.592 (20.007)	-1.312 (4.243)	-1.510 (4.264)	7.389* (4.222)	0.963	0.040**	0.036**	199
Total Number of Teachers (last 6 months)	4.755 (3.244)	-0.375 (0.634)	-0.204 (0.637)	-1.030 (0.631)	0.788	0.299	0.192	199
Average Number of Teachers Working per Day	2.245 (1.762)	-0.005 (0.282)	-0.143 (0.283)	-0.206 (0.280)	0.625	0.472	0.823	199
Chi-Squared Test for Joint Orthogonality of All Variables (p-value)		0.992	0.758	0.052*	0.738	0.043**	0.275	

Notes: .01 - ***; .05 - **; .1 - *. Cross-sectional OLS regressions at CBCC level using baseline data with standard errors (SEs) between parentheses. The last row shows the p-values of a series of Joint Orthogonality Tests: we first estimate a Multinomial Logit where the dependent variable is the lottery group, the explanatory variables are the variables in this table, and the base group is the control group. Then, for column (2), we test the null that the coefficients of explanatory variables that refer to T2 are jointly zero, and analogously for columns (3)-(4). For column (5), we reestimate the Multinomial Logit using T2 as a base group, and test the null that the coefficients that refer to T3 are jointly zero, and analogously for columns (6)-(7).

APPENDIX TABLE 4: Attrition (Primary Caregiver Level)

		Dependent variable: Binary Indicator for Primary Caregiver Lost to Follow-Up			
		18-Month Follow-Up		36-Month Follow-Up	
	Variable	(1)	(2)	(3)	(4)
	T2 (Caregiver training)	-0.023*	-0.026*	-0.007	-0.002
		(0.013)	(0.015)	(0.013)	(0.015)
	T3 (T2 + Caregiver incentives)	0.002	-0.038**	-0.000	-0.013
		(0.016)	(0.016)	(0.014)	(0.015)
	T4 (T2 + Parenting training)	-0.002	-0.025	-0.008	-0.013
		(0.016)	(0.019)	(0.012)	(0.014)
Primary Caregiver Variables	Stimulation Index		-0.009		-0.013
			(0.008)		(0.009)
	Primary Caregiver has a Primary School Leaving Certificate		0.022		-0.006
			(0.026)		(0.018)
	Positive Practices Index		-0.034***		-0.012
			(0.013)		(0.012)
	Strengths and Difficulties Questionnaire: Total Difficulties Score (Inverted)		0.010		-0.009
			(0.013)		(0.013)
	Strengths and Difficulties Questionnaire: Prosocial Score		0.005		0.011
			(0.012)		(0.009)
	Missing Strengths and Difficulties Questionnaire		-0.016		0.018
			(0.044)		(0.023)
	The Center for Epidemiologic Studies Depression Scale		0.022		-0.008
			(0.015)		(0.010)
	Household Wealth Index		0.015		-0.015*
			(0.014)		(0.009)
Interactions: T2 Dummy and Primary Caregiver Variables	T2 x Stimulation Index		0.006		0.017
			(0.011)		(0.013)
	T2 x Primary Caregiver has a Primary School Leaving Certificate		-0.011		0.006
			(0.036)		(0.024)
	T2 x Positive Practices Index		0.033**		0.017
			(0.013)		(0.013)
	T2 x Strengths and Difficulties Questionnaire: Total Difficulties Score (Inverted)		0.004		-0.006
			(0.015)		(0.015)
	T2 x Strengths and Difficulties Questionnaire: Prosocial Score		-0.005		-0.013
			(0.013)		(0.012)
	T2 x Missing Strengths and Difficulties Questionnaire		0.019		-0.055**
			(0.050)		(0.026)
	T2 x The Center for Epidemiologic Studies Depression Scale		-0.012		0.001
			(0.018)		(0.012)
	T2 x Household Wealth Index		0.003		-0.003
			(0.017)		(0.013)

APPENDIX TABLE 4: Attrition (Primary Caregiver Level) - CONTINUED

		18-Month Follow-Up		36-Month Follow-Up	
	Variable	**(1)**	**(2)**	**(3)**	**(4)**
Interactions: T3 Dummy and Primary Caregiver Variables	T3 x Stimulation Index		0.014		0.008
			(0.014)		(0.014)
	T3 x Primary Caregiver has a Primary School Leaving Certificate		0.105*		0.023
			(0.056)		(0.037)
	T3 x Positive Practices Index		0.030*		0.031*
			(0.018)		(0.017)
	T3 x Strengths and Difficulties Questionnaire: Total Difficulties Score (Inverted)		-0.011		-0.002
			(0.019)		(0.017)
	T3 x Strengths and Difficulties Questionnaire: Prosocial Score		-0.022		-0.010
			(0.020)		(0.012)
	T3 x Missing Strengths and Difficulties Questionnaire		0.122*		0.063*
			(0.066)		(0.035)
	T3 x The Center for Epidemiologic Studies Depression Scale		-0.035*		0.013
			(0.020)		(0.015)
	T3 x Household Wealth Index		-0.006		0.008
			(0.022)		(0.011)
Interactions: T4 Dummy and Primary Caregiver Variables	T4 x Stimulation Index		-0.019		-0.003
			(0.029)		(0.013)
	T4 x Primary Caregiver has a Primary School Leaving Certificate		-0.013		-0.013
			(0.037)		(0.028)
	T4 x Positive Practices Index		0.062***		0.027
			(0.019)		(0.018)
	T4 x Strengths and Difficulties Questionnaire: Total Difficulties Score (Inverted)		0.000		0.030*
			(0.020)		(0.015)
	T4 x Strengths and Difficulties Questionnaire: Prosocial Score		-0.004		-0.022*
			(0.016)		(0.012)
	T4 x Missing Strengths and Difficulties Questionnaire		0.077		0.024
			(0.052)		(0.033)
	T4 x The Center for Epidemiologic Studies Depression Scale		0.004		0.031**
			(0.019)		(0.014)
	T4 x Household Wealth Index		0.012		0.021
			(0.029)		(0.014)
	Mean (standard deviation) of the dependent variable for the control group		0.063 (0.243)		0.045 (0.207)
	Joint F-test of Baseline Controls (minus interactions) - p-value		0.181		0.407
	Joint F-test of Interactions (p-value) with T2:		0.362		0.442
	Joint F-test of Interactions (p-value) with T3:		0.045**		0.187
	Joint F-test of Interactions (p-value) with T4:		0.078*		0.127
	Number of observations	2,035	2,035	2,113	2,113

Notes: .01 - ***; .05 - **; .1 - *; Cross-sectional OLS regressions using 18- and 36-month attrition data and baseline variables. Regressions are at the child level, standard errors are clustered by CBCC, and sampling weigths are used (at the 18-month follow-up, 42 randomly tracked observations are reweighted and 85 randomly untracked observations are not included). All regressions control for district-bin fixed effects. All indices are standardized by using (weighted) means and standard deviations from the control group at baseline. The variable Strengths and Difficulties Questionnaire: Total Difficulties Score had its score reversed so that a higher score is better. Missing values in baseline characeristics were replaced with the (weighted) average of that variable for the overall sample at baseline. Following Filmer and Scott (2012), the Household Wealth Index is built using a two-parameter logistic (2PL) Item Response Theory (IRT) model on a set of indicator variables for household assets, infrastructure, and land holdings. The indicator variables include: non-grass roof (e.g. clay, concrete, iron), non-natural floor (e.g. brick, wood), personal source of water (e.g. piped, well), access to electricity, access to a flush toilet, above-median land acreage, and a set of 19 household items such as radio, car and motorcycle.

APPENDIX TABLE 5: Attrition (CBCC Level)

		Dependent variable: Binary Indicator for CBCC Lost to Follow-Up			
		18-Month Follow-Up		36-Month Follow-Up	
	Variable	(1)	(2)	(3)	(4)
	T2 (Caregiver training)	-0.001	-0.047	0.101*	-0.083
		(0.044)	(0.178)	(0.055)	(0.223)
	T3 (T2 + Caregiver incentives)	0.021	-0.246	-0.027	0.009
		(0.044)	(0.201)	(0.055)	(0.252)
	T4 (T2 + Parenting training)	-0.060	-0.161	-0.038	0.015
		(0.044)	(0.152)	(0.055)	(0.191)
CBCC Variables	Classroom Observation Index		0.065*		-0.039
			(0.039)		(0.048)
	Age of the Community-Based Childcare Center (CBCC)		-0.002		0.003
			(0.007)		(0.009)
	Total Number of 3-4 Year-Old Children in the CBCC (observed)		-0.016***		0.009
			(0.006)		(0.007)
	Total Enrollment (reported)		-0.003		0.002
			(0.003)		(0.003)
	Average Daily Attendance (reported)		0.006		-0.005
			(0.004)		(0.005)
	Total Number of Caregivers (last 6 months)		0.006		-0.039**
			(0.013)		(0.016)
	Average Number of Caregivers Working per Day		0.016		0.012
			(0.027)		(0.034)
Interactions: T2 Dummy and CBCC Variables	T2 x Classroom Observation Index		-0.052		0.047
			(0.050)		(0.063)
	T2 x Age of the Community-Based Childcare Center (CBCC)		-0.002		0.006
			(0.013)		(0.016)
	T2 x Total Number of 3-4 Year-Old Children in the CBCC (observed)		0.004		-0.003
			(0.010)		(0.012)
	T2 x Total Enrollment (reported)		-0.001		-0.005
			(0.005)		(0.006)
	T2 x Average Daily Attendance (reported)		0.002		0.008
			(0.007)		(0.009)
	T2 x Total Number of Caregivers (last 6 months)		0.023		0.038
			(0.019)		(0.024)
	T2 x Average Number of Caregivers Working per Day		-0.053		0.009
			(0.040)		(0.050)

APPENDIX TABLE 5: Attrition (CBCC Level) - CONTINUED

	Variable	18-Month Follow-Up (1)	18-Month Follow-Up (2)	36-Month Follow-Up (3)	36-Month Follow-Up (4)
Interactions: T3 Dummy and CBCC Variables	T3 x Classroom Observation Index		-0.103**		0.070
			(0.042)		(0.053)
	T3 x Age of the Community-Based Childcare Center (CBCC)		-0.001		0.009
			(0.013)		(0.017)
	T3 x Total Number of 3-4 Year-Old Children in the CBCC (observed)		0.031***		-0.010
			(0.010)		(0.013)
	T3 x Total Enrollment (reported)		0.006		0.003
			(0.004)		(0.005)
	T3 x Average Daily Attendance (reported)		-0.011*		-0.001
			(0.006)		(0.007)
	T3 x Total Number of Caregivers (last 6 months)		-0.012		0.041*
			(0.017)		(0.021)
	T3 x Average Number of Caregivers Working per Day		-0.057		-0.106
			(0.052)		(0.065)
Interactions: T4 Dummy and CBCC Variables	T4 x Classroom Observation Index		-0.084**		0.005
			(0.042)		(0.053)
	T4 x Age of the Community-Based Childcare Center (CBCC)		-0.002		-0.013
			(0.013)		(0.016)
	T4 x Total Number of 3-4 Year-Old Children in the CBCC (observed)		0.014**		-0.008
			(0.007)		(0.009)
	T4 x Total Enrollment (reported)		0.003		-0.001
			(0.004)		(0.005)
	T4 x Average Daily Attendance (reported)		-0.007		0.002
			(0.006)		(0.007)
	T4 x Total Number of Caregivers (last 6 months)		-0.007		-0.024
			(0.020)		(0.025)
	T4 x Average Number of Caregivers Working per Day		0.003		0.092
			(0.047)		(0.059)
	Mean (stardard deviation) of dependent variable for the control group	0.061 (0.242)		0.102 (0.306)	
	Joint F-test of Baseline Controls (minus interactions) - p-value		0.188		0.343
	Joint F-test of Interactions - p-value with T2:		0.613		0.718
	Joint F-test of Interactions - p-value with T3:		0.025**		0.359
	Joint F-test of Interactions - p-value with T4:		0.246		0.743
	Number of observations	199	199	199	199

Notes: .01 - ***; .05 - **; .1 - *; Cross-sectional OLS regressions using 18- and 36-month attrition data and baseline variables. All regressions control for district-bin fixed effects. Classroom Observation Index is obtained by weighting the underlying Classroom Observation variables using Inverse Covariance Weights (ICW), as in Casey et al. (2012) and described in detail in the Appendix.

APPENDIX TABLE 6: Impacts on Child's Behavioral Problems - 18-Month and 36-Month Follow-Ups

PANEL A: 18-Month Follow-Up		**Dependent variable: Strengths and Difficulties Questionnaire Subscale Scores**							
		Emotion (Inverted)		**Conduct (Inverted)**		**Hyperactivity (Inverted)**		**Peer Problems (Inverted)**	
Variables		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
T2 (caregiver training)		-0.058	-0.071	-0.006	-0.019	-0.031	-0.059	-0.009	-0.006
		(0.069)	(0.065)	(0.055)	(0.052)	(0.065)	(0.063)	(0.075)	(0.072)
T3 (T2 + caregiver incentives)		0.044	0.030	0.084	0.079	0.106*	0.095	0.033	0.034
		(0.063)	(0.062)	(0.056)	(0.056)	(0.062)	(0.061)	(0.076)	(0.076)
T4 (T2 + parenting training)		0.008	-0.018	0.084*	0.059	0.152**	0.128**	0.057	0.048
		(0.062)	(0.059)	(0.050)	(0.050)	(0.060)	(0.056)	(0.077)	(0.074)
Lagged Dependent Variable (Baseline)			0.249***		0.340***		0.162***		0.143***
			(0.025)		(0.023)		(0.026)		(0.030)
Missing Lagged Dependent Variable (Baseline)			-0.103		-0.039		0.042		-0.063
			(0.127)		(0.090)		(0.082)		(0.136)
F-test for Equality of Parameters - p-value	T2=T3	0.120	0.116	0.106	0.080*	0.025**	0.012**	0.545	0.561
	T2=T4	0.298	0.376	0.076*	0.111	0.003***	0.002***	0.343	0.430
	T3=T4	0.522	0.391	0.993	0.709	0.443	0.567	0.741	0.843
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes	No	Yes	No	Yes
Number of observations		1,938	1,938	1,938	1,938	1,938	1,938	1,938	1,938

APPENDIX TABLE 6: Impacts on Child's Behavioral Problems - 18-Month and 36-Month Follow-Ups (CONTINUED)

PANEL B: 36-Month Follow-Up		**Dependent variable: Strengths and Difficulties Questionnaire Subscale Scores**							
		Emotion (Inverted)		**Conduct (Inverted)**		**Hyperactivity (Inverted)**		**Peer Problems (Inverted)**	
Variables		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
T2 (caregiver training)		0.169***	0.150***	0.039	0.030	0.075	0.060	0.059	0.043
		(0.054)	(0.052)	(0.057)	(0.055)	(0.064)	(0.062)	(0.069)	(0.070)
T3 (T2 + caregiver incentives)		0.113	0.120*	0.051	0.062	0.075	0.092	-0.028	-0.013
		(0.069)	(0.068)	(0.060)	(0.059)	(0.061)	(0.059)	(0.075)	(0.075)
T4 (T2 + parenting training)		-0.007	-0.010	-0.071	-0.074	0.002	0.004	-0.064	-0.065
		(0.061)	(0.059)	(0.057)	(0.055)	(0.066)	(0.067)	(0.065)	(0.064)
Lagged Dependent Variable (Baseline)			0.163***		0.252***		0.141***		0.145***
			(0.023)		(0.024)		(0.025)		(0.025)
Missing Lagged Dependent Variable (Baseline)			0.071		0.005		0.180**		0.071
			(0.113)		(0.082)		(0.084)		(0.098)
F-test for Equality of Parameters - p-value	T2=T3	0.350	0.625	0.832	0.580	0.999	0.577	0.252	0.478
	T2=T4	0.001***	0.002***	0.055*	0.048**	0.217	0.349	0.065*	0.114
	T3=T4	0.082*	0.053*	0.038**	0.016**	0.184	0.132	0.612	0.464
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Lagged Dependent Variable and Age Dummies?		No	Yes	No	Yes	No	Yes	No	Yes
Number of observations		2,022	2,022	2,022	2,022	2,022	2,022	2,022	2,022

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at child level using data on the Strengths and Difficulties Questionnaire. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the relevant variables are available. Each subscale from the Strengths and Difficulties Questionnaire had its score inverted so that a higher score is better. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up). For each round, indices are stardardized using the (weighted) mean and standard deviation for the control group in that round.

APPENDIX TABLE 7: Impacts on Child Stimulation Activities Performed by the Primary Caregiver

PANEL A: 18-Month Follow-Up

Dependent variable: Child Stimulation Activities Performed by the Primary Caregiver	Explanatory variables				Mean (standard deviation) of dependent variable in the control group	F-test for Equality of Parameters (p-values)			Number of observations
	T2 (caregiver training)	T3 (T2 + caregiver incentives)	T4 (T2 + parenting training)	Lagged Dependent Variable (Baseline)		T2=T3	T2=T4	T3=T4	
(1) Read Books or Looked at Pictures	0.046 (0.034)	0.028 (0.038)	0.090** (0.037)	0.185*** (0.029)	0.309 (0.463)	0.559	0.132	0.046**	1,937
(2) Told Stories	-0.074** (0.033)	-0.009 (0.033)	0.026 (0.033)	0.081*** (0.028)	0.652 (0.477)	0.037**	0.002***	0.262	1,937
(3) Sang a Song	-0.003 (0.030)	-0.001 (0.031)	0.009 (0.031)	0.114*** (0.030)	0.609 (0.488)	0.958	0.680	0.745	1,937
(4) Chatted while Doing Chores	-0.017 (0.033)	-0.026 (0.036)	0.057* (0.032)	0.025 (0.024)	0.657 (0.475)	0.786	0.008***	0.009***	1,937
(5) Took Outside the Home	-0.074** (0.036)	-0.024 (0.034)	0.042 (0.030)	0.018 (0.032)	0.341 (0.475)	0.186	0.000***	0.038**	1,937
(6) Played at Physical Activities	0.019 (0.036)	0.027 (0.041)	0.118*** (0.034)	0.064** (0.026)	0.409 (0.492)	0.840	0.002***	0.012**	1,937
(7) Helped Learn Letters or Numbers	0.044 (0.033)	0.020 (0.033)	0.163*** (0.036)	0.078*** (0.028)	0.394 (0.489)	0.410	0.000***	0.000***	1,937
(8) Helped Learn Shapes or Colors	0.028 (0.027)	0.083** (0.034)	0.151*** (0.028)	0.053 (0.042)	0.210 (0.408)	0.118	0.000***	0.059*	1,937
(9) Drew Objects in Sand or Paper	0.023 (0.032)	0.023 (0.033)	0.093*** (0.031)	0.114*** (0.031)	0.362 (0.481)	0.995	0.023**	0.034**	1,937
(10) Construed Objects (paper, wire, mud, etc.)	-0.064* (0.034)	-0.019 (0.035)	-0.014 (0.037)	0.029 (0.045)	0.346 (0.476)	0.124	0.087*	0.857	1,937
(11) Identified Plants or Animals	-0.018 (0.031)	0.056* (0.032)	0.052* (0.030)	0.042 (0.034)	0.225 (0.418)	0.025**	0.020**	0.891	1,937
(12) Taught English Words	0.023 (0.027)	0.010 (0.027)	0.066** (0.027)	0.243*** (0.033)	0.210 (0.408)	0.604	0.098*	0.036**	1,937
(13) Taught Names and Uses of New Objects	-0.048* (0.025)	-0.012 (0.027)	0.043 (0.026)	0.054 (0.034)	0.201 (0.402)	0.148	0.000***	0.045**	1,937

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using baseline and 18-month follow-up data on Stimulation. Standard errors in parentheses are clustered at the CBCC level. All regressions control for district-bin fixed effects and age dummies. The variable referred to as 'baseline control variable' is the lagged (baseline) value of the dependent variable. We use all children for whom the dependent variable is available for the 18-month follow-up. We weight observations using sampling weights (and tracking weights for the 42 randomly tracked observations at the 18-month follow-up).

APPENDIX TABLE 7: Impacts on Child Stimulation Activities Performed by the Primary Caregiver (CONTINUED)

PANEL B: 36-Month Follow-Up

Dependent variable: Child Stimulation Activities Performed by the Primary Caregiver	Explanatory variables					Mean (standard deviation) of dependent variable in the control group	F-test for Equality of Parameters (p-values)			Number of observations
	T2 (caregiver training)	T3 (T2 + caregiver incentives)	T4 (T2 + parenting training)	Lagged Dependent Variable (Baseline)	Lagged Stimulation Index (Baseline)		T2=T3	T2=T4	T3=T4	
(1) Read Books or Looked at Story Books	0.028	(0.048)	(0.074)	0.123		0.370	0.481	0.079*	0.355	2,023
	(0.028)	0.030**	(0.028)	(0.030)		(0.483)				
(2) Told Stories	0.053*	0.043	0.057*	0.048**		0.569	0.739	0.885	0.634	2,023
	(0.032)	(0.031)	(0.032)	(0.024)		(0.496)				
(3) Sang a Song	0.016	-0.014	0.001	0.100***		0.564	0.335	0.614	0.630	2,023
	(0.031)	(0.033)	(0.031)	(0.030)		(0.496)				
(4) Played Sports	0.014	0.040	0.041	0.016		0.450	0.409	0.334	0.968	2,023
	(0.027)	(0.031)	(0.029)	(0.025)		(0.498)				
(5) Named or Recognized Letters or Numbers	-0.055	0.032	0.059	0.077***		0.490	0.014**	0.001***	0.439	2,023
	(0.039)	(0.038)	(0.037)	(0.027)		(0.500)				
(6) Helped Learn Shapes or Colors	-0.063**	-0.029	0.012	0.031		0.320	0.256	0.004***	0.155	2,023
	(0.030)	(0.032)	(0.027)	(0.038)		(0.467)				
(7) Drew Objects in Sand or Paper	-0.010	0.040	0.064*	0.064**		0.344	0.145	0.037**	0.539	2,023
	(0.031)	(0.035)	(0.037)	(0.030)		(0.475)				
(8) Construed Objects (paper, wire, mud, etc.)	-0.019	-0.010	0.026	0.049		0.362	0.792	0.157	0.241	2,023
	(0.037)	(0.035)	(0.031)	(0.036)		(0.481)				
(9) Identified Plants or Animals	0.018	-0.033	0.045	0.029		0.304	0.125	0.382	0.018**	2,023
	(0.036)	(0.036)	(0.038)	(0.029)		(0.460)				
(10) Taught English Words	-0.039	0.003	0.035	0.242***		0.279	0.113	0.003***	0.293	2,023
	(0.029)	(0.033)	(0.034)	(0.036)		(0.449)				
(11) Wrote Letters, Numbers or Words	-0.111***	0.011	0.019		0.058***	0.505	0.000***	0.000***	0.792	2,023
	(0.039)	(0.037)	(0.038)		(0.015)	(0.500)				
(12) Helped with Homework	-0.061*	0.001	0.030		0.059***	0.307	0.021**	0.001***	0.305	2,023
	(0.031)	(0.031)	(0.031)		(0.013)	(0.462)				
(13) Taugh Words in Local Language	0.009	0.037	0.066*		0.046***	0.305	0.372	0.088*	0.432	2,023
	(0.036)	(0.037)	(0.040)		(0.015)	(0.461)				
(14) Talked About School Content	-0.037	0.015	0.029		0.056***	0.643	0.131	0.050**	0.684	2,023
	(0.035)	(0.036)	(0.032)		(0.013)	(0.480)				
(15) Took on a Fun Outing (football, dance, etc.)	-0.027	0.009	0.031		0.016	0.410	0.298	0.112	0.571	2,023
	(0.033)	(0.036)	(0.035)		(0.015)	(0.492)				
(16) Did Dances	0.040	0.045	0.034		0.043***	0.327	0.863	0.851	0.722	2,023
	(0.029)	(0.030)	(0.027)		(0.013)	(0.469)				
(17) Played Music Instruments	0.025	0.017	0.043**		0.023*	0.146	0.736	0.441	0.269	2,023
	(0.021)	(0.023)	(0.021)		(0.012)	(0.354)				
(18) Took on Errands (market, etc.)	0.015	0.023	0.050		0.023	0.475	0.799	0.338	0.416	2,026
	(0.036)	(0.032)	(0.031)		(0.017)	(0.500)				

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using baseline and 36-month follow-up data on Stimulation. Standard errors in parentheses are clustered at the CBCC level. All regressions control for district-bin fixed effects and age dummies. We use all children for whom the dependent variable is available for the 36-month follow-up. The variable referred to as 'baseline control variable' is the lagged (baseline) value of the dependent variable, except in rows (11)-(18), where it is the baseline value of the Stimulation Index. This is because these activities were only asked at the 36-month follow-up. We weight observations using sampling weights.

APPENDIX TABLE 8: Impacts on Classroom Quality

PANEL A: 18-Month Follow-Up		**Dependent variable: Classroom Observation Sub-indices**			
		ICW Index of Subscales		**Principal Components Analysis**	
		Structure, Engagement, Supervision	**Literacy, Numeracy, Problem Solving, Motor Activities**	**First PC (Supervision and Engagement)**	**Second PC (Counting, Teaching Alphabet, Fine Motor)**
Variables		**(1)**	**(2)**	**(3)**	**(4)**
T2 (caregiver training)		0.609**	-0.014	0.449**	0.121
		(0.244)	(0.235)	(0.196)	(0.195)
T3 (T2 + caregiver incentives)		0.959***	0.559**	0.721***	0.519***
		(0.246)	(0.242)	(0.199)	(0.198)
T4 (T2 + parenting training)		0.758***	0.667***	0.660***	0.572***
		(0.237)	(0.234)	(0.192)	(0.191)
Lagged Dependent Variable (Baseline)		0.050	-0.016	-0.001	-0.056
		(0.071)	(0.061)	(0.082)	(0.078)
F-test for Equality of Parameters (p-value)	T2=T3	0.156	0.018**	0.173	0.045**
	T2=T4	0.531	0.004***	0.272	0.019**
	T3=T4	0.402	0.640	0.754	0.783
District-bin Fixed Effects?		Yes	Yes	Yes	Yes
Lagged Dependent Variable?		Yes	Yes	Yes	Yes
Number of observations		189	189	189	189

APPENDIX TABLE 8: Impacts on Classroom Quality (CONTINUED)

PANEL B: 36-Month Follow-Up		**Dependent variable: Classroom Observation Sub-indices**			
		ICW Index of Subscales		**Principal Components Analysis**	
		Structure, Engagement, Supervision	**Literacy, Numeracy, Problem Solving, Motor Activities**	**First PC (Supervision and Engagement)**	**Second PC (Fine Motor Activities, Identifying Shapes/Colors, Grouping Children)**
Variables		**(1)**	**(2)**	**(3)**	**(4)**
T2 (caregiver training)		0.556	0.245	0.243	0.527**
		(0.351)	(0.265)	(0.249)	(0.247)
T3 (T2 + caregiver incentives)		0.464	0.260	0.455*	0.257
		(0.333)	(0.257)	(0.237)	(0.238)
T4 (T2 + parenting training)		0.320	0.280	0.221	0.513**
		(0.334)	(0.261)	(0.240)	(0.239)
Lagged Dependent Variable (Baseline)		0.065	-0.011	0.147	-0.010
		(0.102)	(0.070)	(0.104)	(0.097)
F-test for Equality of Parameters (p-value)	T2=T3	0.792	0.957	0.388	0.274
	T2=T4	0.503	0.899	0.932	0.953
	T3=T4	0.662	0.936	0.325	0.275
District-bin Fixed Effects?		Yes	Yes	Yes	Yes
Lagged Dependent Variable?		Yes	Yes	Yes	Yes
Number of observations		178	178	178	178

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at CBCC level using data on Classroom Observations. Standard errors in parentheses. For each round, we use all non-missing CBCCs. Subcomponent indices (columns (1)-(2)) are obtained by weighting underlying variables using Inverse Covariance Weights (ICW), as in Casey et al. (2012) and described in detail in the Appendix. Columns (3) and (4) present the first two principal components (PC) obtained by running Principal Component Analysis on the underlying variables in the Classroom Observation tool. For each round, indices are stardardized using the mean and standard deviation for the control group in that round.

APPENDIX TABLE 9: Impacts on Child Assessments - 36-Month Follow-Up

	Kaufman Assessment Battery for Children Scores						Early Grade Math Assessment Scores					
	Hand Movements		Number Recall		Triangles		Number Recognition		Quantity Discrimination		Addition	
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**
T2 (caregiver training)	-0.027	-0.011	0.024	0.042	0.037	0.051	-0.153**	-0.134**	-0.079	-0.060	-0.064	-0.051
	(0.077)	(0.080)	(0.077)	(0.076)	(0.083)	(0.079)	(0.069)	(0.060)	(0.074)	(0.062)	(0.079)	(0.073)
T3 (T2 + caregiver incentives)	0.003	0.003	-0.075	-0.075	0.165**	0.154**	-0.070	-0.077	-0.040	-0.043	0.010	0.013
	(0.082)	(0.084)	(0.081)	(0.081)	(0.076)	(0.076)	(0.071)	(0.061)	(0.072)	(0.063)	(0.081)	(0.077)
T4 (T2 + parenting training)	0.025	0.052	-0.028	0.004	0.096	0.120	-0.030	-0.011	0.014	0.032	-0.082	-0.064
	(0.085)	(0.084)	(0.079)	(0.077)	(0.078)	(0.076)	(0.074)	(0.063)	(0.077)	(0.067)	(0.072)	(0.068)
Malawi Developmental Assessment Tool: Fine Motor / Perception Skills (Baseline)		0.230***		0.265***		0.374***		0.367***		0.372***		0.331***
		(0.047)		(0.032)		(0.036)		(0.035)		(0.035)		(0.042)
F-test for Equality of Parameters - p-value T2=T3	0.684	0.850	0.194	0.124	0.114	0.182	0.195	0.325	0.555	0.782	0.396	0.438
F-test for Equality of Parameters - p-value T2=T4	0.470	0.395	0.466	0.595	0.459	0.354	0.064*	0.050**	0.166	0.143	0.823	0.861
F-test for Equality of Parameters - p-value T3=T4	0.772	0.521	0.528	0.282	0.350	0.646	0.541	0.263	0.397	0.216	0.260	0.333
District-bin Fixed Effects?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Control Variable and Age Dummies?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Number of observations	2,028	2,028	2,027	2,027	2,028	2,028	2,027	2,027	2,025	2,025	2,017	2,017

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the child level using baseline and 36-month data on scores. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the relevant variables are available. We weight observations using sampling weights. For each round, scores are stardardized using the (weighted) mean and standard deviation for the control group in that round.

APPENDIX TABLE 10: Correlation Coefficients of Child Assessments

		Baseline				18-Month Follow-Up	
		Malawi Developmental Assessment Tool: Fine Motor / Perception Skills	Malawi Developmental Assessment Tool: Language Skills	Peabody Picture Vocabulary Test	Leiter Sustained Attention	Malawi Developmental Assessment Tool: Fine Motor / Perception Skills	Malawi Developmental Assessment Tool: Language Skills
Baseline	Malawi Developmental Assessment Tool: Fine Motor / Perception Skills	1.000					
	Malawi Developmental Assessment Tool: Language Skills	0.682***	1.000				
	Peabody Picture Vocabulary Test	0.584***	0.568***	1.000			
	Leiter Sustained Attention	0.200***	0.166**	0.226***	1.000		
18-Month Follow-Up	Malawi Developmental Assessment Tool: Fine Motor / Perception Skills	0.363***	0.369***	0.263***	0.153**	1.000	
	Malawi Developmental Assessment Tool: Language Skills	0.316***	0.433***	0.214***	0.044	0.657***	1.000
36-Month Follow-Up	Peabody Picture Vocabulary Test	0.185***	**0.235*** **	**0.237*** **	0.069	0.282***	**0.295*** **
	Leiter Sustained Attention	0.309***	0.309***	0.357***	0.108	0.431***	0.387***
	Kaufman Assessment Battery for Children Score	0.240***	0.294***	0.271***	0.155**	0.403***	0.370***
	Early Grade Math Assessment Score	0.349***	0.413***	0.419***	0.146**	0.486***	0.443***

Notes: This table presents bivariate correlation coefficients for average child assessment scores at the CBCC level. For each round and assessment, the scores of individual children at a CBCC are averaged using sampling weights. We use raw scores for every assessment, with exception of Kaufman Assessment Battery for Children and Early Grade Math Assessment. For each of these two assessments, total scores were obtained by constructing a weighted average of the subscale scores of each assessment (for Kaufman Assessment Battery for Children: Hand Movements, Triangles, and Number Recall / for Early Grade Math Assessment: Number Recognition, Quantity Discrimination, and Addition), where the weights were obtained using Inverse Covariance Weighting, as in Casey et al. (2012) and described in detail in the Appendix.

APPENDIX TABLE 11: Impacts on Child Schooling Outcomes - 36-Month Follow-Up

		Dependent Variable:				
		Difference Between Actual Grade and Age-Appropriate Grade	Share of Class Days Attended in the Previous Week	Primary Caregiver Met Teacher to Discuss Progress	Primary Caregiver Met Teacher to Discuss Behavior	Child Repeating a Grade
Variables		**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
T2 (caregiver training)		-0.120**	0.004	-0.045	-0.023	0.005
		(0.049)	(0.022)	(0.028)	(0.031)	(0.028)
T3 (T2 + caregiver incentives)		-0.056	0.009	-0.010	-0.031	-0.021
		(0.045)	(0.022)	(0.026)	(0.025)	(0.028)
T4 (T2 + parenting training)		-0.046	0.008	0.039	0.030	0.002
		(0.046)	(0.021)	(0.027)	(0.028)	(0.027)
Mean (standard deviation) for dependent variable in the control group		-0.214	0.830	0.227	0.189	0.305
		(0.749)	(0.268)	(0.419)	(0.392)	(0.461)
F-test for Equality of Parameters - p-value	T2=T3	0.214	0.789	0.210	0.778	0.363
	T2=T4	0.147	0.826	0.003***	0.085*	0.938
	T3=T4	0.830	0.953	0.079*	0.026**	0.408
District-bin Fixed Effects?		Yes	Yes	Yes	Yes	Yes
Age Dummies?		Yes	Yes	Yes	Yes	Yes
Number of observations		1,999	1,991	1,999	1,999	1,996

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at child level using 36-month data on Child Schooling Variables. Standard errors in parentheses are clustered at the CBCC level. We use all children for whom the dependent variable is available for the 36-Month Follow-Up. For the dependent variable in column (1), we assume that the appropriate grade for children aged less than 84 months is grade 1, for children aged between 84 and 95 months is grade 2, and for children aged above 95 months old is grade 3. We weight observations using our sampling weights.

APPENDIX TABLE 12: Impacts on Teacher Characteristics (CBCC Level)

PANEL A: 18-Month Follow-Up	Dependent variable:				
	Number of Active Teachers with Education Info	**Number of Active Teachers with a Primary School Leaving Certificate**	**Share of Active Teachers with Primary School Leaving Certificate**	**Average Teacher Age**	**Share of Active Teachers with at Least Two Years of Tenure**
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
T2 (caregiver training)	-0.064	0.199	0.118**	-2.883**	-0.011
	(0.318)	(0.257)	(0.053)	(1.258)	(0.069)
T3 (T2 + caregiver incentives)	-0.502	0.105	0.182***	-1.793	0.000
	(0.323)	(0.260)	(0.054)	(1.266)	(0.070)
T4 (T2 + parenting training)	-0.533*	-0.072	0.131**	-2.468**	-0.090
	(0.308)	(0.247)	(0.051)	(1.216)	(0.068)
Lagged Dependent Variable (Baseline)	0.752***	0.210*	0.335***	0.600***	0.087
	(0.131)	(0.115)	(0.063)	(0.056)	(0.065)
Mean (and standard deviation) of dependent variable for the control group	3.239	2.152	0.691	36.136	0.711
	(1.946)	(1.673)	(0.368)	(8.885)	(0.359)
F-test for Equality of Parameters - p-value T2=T3	0.174	0.715	0.238	0.387	0.875
F-test for Equality of Parameters - p-value T2=T4	0.131	0.273	0.799	0.732	0.236
F-test for Equality of Parameters - p-value T3=T4	0.922	0.482	0.336	0.583	0.185
Number of observations	188	188	188	189	189

APPENDIX TABLE 12: Impacts on Teacher Characteristics (CBCC Level) - CONTINUED

PANEL B: 36-Month Follow-Up	Dependent variable:				
	Number of Active Teachers with Education Info	Number of Active Teachers with a Primary School Leaving Certificate	Share of Active Teachers with Primary School Leaving Certificate	Average Teacher Age	Share of Active Teachers with at Least Two Years of Tenure
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**
T2 (caregiver training)	-0.357	0.095	0.142**	-2.322	-0.022
	(0.217)	(0.180)	(0.056)	(1.689)	(0.080)
T3 (T2 + caregiver incentives)	-0.250	0.164	0.143***	-1.429	-0.063
	(0.208)	(0.172)	(0.053)	(1.603)	(0.076)
T4 (T2 + parenting training)	-0.305	0.052	0.087	-0.261	-0.071
	(0.210)	(0.173)	(0.053)	(1.603)	(0.077)
Lagged Dependent Variable (Baseline)	0.225**	0.012	0.269***	0.666***	0.112
	(0.087)	(0.078)	(0.065)	(0.077)	(0.074)
Mean (and standard deviation) of dependent variable for the control group	2.386	1.705	0.752	36.032	0.742
	(1.104)	(0.978)	(0.316)	(8.805)	(0.306)
F-test for Equality of Parameters - p-value T2=T3	0.618	0.697	0.989	0.589	0.608
F-test for Equality of Parameters - p-value T2=T4	0.811	0.814	0.319	0.219	0.544
F-test for Equality of Parameters - p-value T3=T4	0.790	0.512	0.290	0.461	0.918
Number of observations	178	178	178	178	178

Notes: .01 - ***; .05 - **; .1 - *; OLS regressions at the CBCC level include all CBCCs for which relevant variables are available. District-bin fixed effects are included.