I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11212

Preschool Attendance, School Progression, and Cognitive Skills in East Africa

Jan Bietenbeck
Sanna Ericsson
Fredrick M. Wamalwa

DECEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11212

Preschool Attendance, School Progression, and Cognitive Skills in East Africa

Jan Bietenbeck
Lund University, DIW Berlin and IZA

Sanna Ericsson
Lund University

Fredrick M. Wamalwa
University of Cape Town

DECEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Preschool Attendance, School Progression, and Cognitive Skills in East Africa*

We study the effects of preschool attendance on children's school progression and cognitive skills in Kenya and Tanzania. Our analysis uses novel data from large-scale household surveys of children's literacy and numeracy skills, which also collect retrospective information on preschool attendance. Against the backdrop of a large expansion of pre-primary education, our regressions identify the impacts from within-household differences, controlling for a variety of child-specific covariates. In both countries, children who go to preschool tend to enroll in primary school late, and thus fall behind in terms of grades completed at early ages. However, once in school, they progress through grades faster and at ages 13-16 have completed about one and a half more months of schooling than their same-aged peers who did not attend preschool. They also score around 0.10 standard deviations higher on standardized cognitive tests, showing that there are important longer-term benefits from preschool in Kenya and Tanzania.

JEL Classification: I21, J24
Keywords: preschool, education, cognitive skills, Sub-Saharan Africa

Corresponding author:
Jan Bietenbeck
Department of Economics
Lund University
P.O. Box 7082
220 07 Lund
Sweden
E-mail: mail@janbietenbeck.com

* We thank Kaveh Majlesi, Alessandro Martinello, Andreas Stegmann, Alessandro Tarozzi, Petra Thiemann, and audiences at the Lund-SFI Workshop in Applied Microeconomics, the Nordic Conference in Development Economics, and the Copenhagen Education Network Workshop for helpful comments.

1. Introduction

School enrollment in Sub-Saharan Africa has increased substantially over the past two decades. However, many students fall behind the curriculum early on, and grade repetition and early dropout are widespread (UNESCO, 2012). Students also learn remarkably little in school: for example, only one in five third-grade students in East Africa has second-grade literacy and numeracy skills, and less than one third of sixth-grade students in Southern and Eastern Africa can solve a simple subtraction problem (Uwezo, 2015; Bietenbeck, Piopiunik, and Wiederhold, 2017).

One possible reason why students in these countries perform so poorly is that they enter school unprepared. Specifically, to the extent that early and later learning are complementary (Cunha and Heckman, 2007), a lack of education before starting school reduces children's efficiency in learning once they arrive there. High-quality preschool programs which prepare children for school are therefore often seen as a promising way to enhance learning outcomes (e.g. World Bank, 2018). In Sub-Saharan Africa, pre-primary education has been expanding rapidly, with the gross enrollment ratio doubling from 11% to 22% between 2000 and 2014.[1] Whether preschools are actually effective at boosting student outcomes in this context is unclear, however, because rigorous empirical evidence is still scarce.

In this paper, we study the effects of preschool attendance on children's school progression and cognitive skills in Kenya and Tanzania. Our empirical analysis draws on data from Uwezo, which conducts nationally representative household surveys of school-age children's education and their literacy and numeracy skills. The surveys also collect retrospective information on preschool attendance, which we can relate to current outcomes of respondents up to 16 years of age. Our investigation focuses on two main outcomes: the highest grade of school completed and a cognitive test score, which summarizes a child's performance on the standardized literacy and numeracy tests. The data contain information on these measures for more than half a million children across the two countries, independently of whether they are currently enrolled in school or not.

[1] These figures come from the World Bank's World Development Indicators and can be accessed here: https://data.worldbank.org/indicator/SE.PRE.ENRR.

Our regressions compare the outcomes of children of the same age who did and did not attend preschool. For identification, we rely on within-household differences, thereby controlling for all determinants of outcomes and attendance that vary across families.[2] We argue that the leftover variation between siblings is likely due to changes in the local availability of preschools, which came about because of a large expansion of the pre-primary sector during our study period. In support of this claim, we show that even within households, children in later cohorts are much more likely to have attended preschool. To mitigate any remaining concerns about endogenous selection, the regressions control for a variety of predetermined characteristics that vary between siblings. Finally, we also show that children who attended preschool do not differentially benefit from other educational inputs such as private tutoring, suggesting that our estimates are not driven by unobserved child-specific investments.

The impact of preschool attendance on school progression follows an interesting dynamic pattern. In both Kenya and Tanzania, children often enroll in preschool late and only proceed to primary school once they finished it. At early ages, these children have therefore completed fewer school grades than their same-aged peers who did not attend preschool. However, we find that once enrolled in primary school, children who attended preschool progress through grades faster and eventually overtake their peers. Thus, at ages 13-16, they have accumulated about one and a half more months of schooling in both countries.

The results for cognitive test scores similarly show that children who went to preschool outperform their peers in the long run. In Kenya, this effect fades in at early ages and soon stabilizes at a gain of 0.12 standard deviations (SD). In Tanzania, in contrast, an initial gain of 0.11 SD fades out slightly to 0.08 SD at ages 13-16. All these estimates are robust to a variety of sensitivity checks, and we show that they are unlikely to be driven by systematic recall error in the retrospectively reported preschool variable. Overall, our findings demonstrate that there are important longer-

[2]Similar strategies have been used by Currie and Thomas (1995), Garces, Thomas, and Currie (2002), and Deming (2009) to estimate the impacts of the Head Start program in the United States, and by Berlinski, Galiani, and Manacorda (2008) to estimate the effects of preschool in Uruguay.

term gains from preschool attendance in Kenya and Tanzania.

Our paper contributes to a small literature on the impacts of preschool education on children's outcomes in developing countries, which has mostly focused on Latin America and Asia. Using retrospective data on preschool enrollment and the same within-household estimator as we do, Berlinski, Galiani, and Manacorda (2008) find that Uruguayan children who attended preschool accumulate 0.8 more years of education by age 15. Similarly, Behrman, Cheng, and Todd (2004) and Berlinski, Galiani, and Gertler (2009) document positive short-term effects of preschool attendance on children's cognitive skills in Bolivia and Argentina, respectively. In contrast, a randomized evaluation of a preschool construction program in Cambodia found negative short-term impacts on test scores of targeted children, a result that is partly explained by a shift from underage enrollment in primary school to enrollment in preschool (Bouguen et al., 2017).

To the best of our knowledge, the only other rigorous study of preschool effects in Sub-Saharan Africa is the paper by Martinez, Naudeau, and Pereira (2013), who report on an experimental evaluation of a model preschool program in the Gaza province of Mozambique. The authors find that two years after the start of the program, children were more likely to be enrolled in primary school and had higher cognitive and socio-emotional skills. In contrast to this small-scale evaluation, our study uses nationally representative data on preschool attendance and learning outcomes from two countries. Moreover, unlike most of the previous literature, we are able to examine the longer-term effects of preschool attendance.

2. Institutional background

2.1. Preschools in Kenya

Pre-primary education in Kenya comprises three grades: baby class (ages 3-4), nursery (ages 4-5), and pre-unit (ages 5-6). Attendance is not compulsory, but a large and rising share of children does enroll in preschool. For example, whereas 79% of the cohort born in 1997 attended preschool,

the corresponding figure for the 2004 cohort is 83%.[3] In practice, children often start pre-primary education late, with some enrolling even after they have reached the official school entry age of 6 years. This behavior is explained by the fact that many parents see preschool as an integral part of children's education, and therefore prefer them to complete it before proceeding to primary school (Bidwell, Parry, and Watine, 2013). Finally, children who do not go to preschool typically stay home to help with household chores, with some five-year olds enrolling in primary school early.

There are two broad types of preschools in Kenya. First, public preschools are run by the government and are usually attached to a primary school. In rural areas, these preschools are often the only available option. Second, private preschools are owned and run by a variety of providers, including non-governmental and faith-based organizations, community-based associations, and private-for-profit agents. They comprise a diverse range of institutions, including: (a) highly unregulated and unregistered non-formal preschools, which are mainly located in informal urban settlements; (b) formal private preschool academies in middle and high-income areas; and (c) a small number of exclusive private preschools catering to very high-income households. All preschools (both public and private) charge tuition fees, which vary widely depending on the type of institution attended, with non-formal private preschools charging the lowest fees.[4]

Unlike in many European countries and the United States, where pre-primary education for younger children typically focuses on play, preschool studies in Kenya are highly academic: students sit at desks and listen to the teacher teach in a classroom-like setting. Curricula, while not standardized, tend to emphasize the learning of basic numeracy and literacy skills via memorization and recitation. In contrast, only little attention is paid to the development of socio-emotional skills (Bidwell, Parry, and Watine, 2013).

[3] All enrollment statistics in this section are based on data from the Uwezo surveys, which we describe in detail below. Here, we focus on cohorts observed when they are at least 10 years old in order to account for late enrollment in preschool.

[4] For households in informal urban settlements, non-formal private pre-schools are sometimes the only option available (Piper, Zuilkowski, and Mugenda, 2014). For overviews of the institutional landscape in Kenyan education, see Tooley, Dixon, and Stanfield (2008), Heyneman and Stern (2014), and Edwards Jr., Klees, and Wildish (2015).

2.2. Preschool in Tanzania

In Tanzania, children attend preschool for up to two years before entering primary school at age 7, although late enrollment is quite common. Preschool is not compulsory but increasingly popular, with the attendance rate rising from 61% to 69% for the cohorts born in 1995 and 2004, respectively. This rise is spurred by the Tanzanian government, whose expansion strategy has been to attach pre-primary classrooms to existing primary schools. Due to limited funding, however, most of these classroom are under-resourced and overcrowded, with the average student-to-teacher ratio being 100:1 in 2011 (World Bank, 2012).

Unlike in Kenya, private preschools serve only a very small fraction of children in Tanzania. Public preschools are subsidized by the government, but they do charge varying tuition fees. In contrast to the situation in Kenya, there is an official government preschool curriculum, which emphasizes the development of both cognitive and socio-emotional skills. In practice, however, preschool teachers often have little or no knowledge of the official curriculum and tend to focus on formal instruction in basic literacy and numeracy (Mligo, 2016).

3. Data

3.1. The Uwezo surveys

The Uwezo initiative has been conducting large-scale assessments of school-age children's literacy and numeracy skills in Kenya, Tanzania, and Uganda since 2009. The assessments are administered as part of repeated cross-sectional household surveys, which are representative at the district level. An important advantage of this design is that skills are measured also for children who are currently not enrolled in school. The surveys collect information from children aged 6-16 (7-16 in Tanzania) on their current enrollment and highest grade completed as well as on a variety of child and household characteristics. Crucially for our purposes, in recent waves, respondents were also asked whether they ever attended preschool.[5]

[5]In Uganda, the only nationally representative Uwezo survey which asked about preschool attendance was conducted in 2013. Unfortunately, information on this key variable is missing for 49% of the respondents in this wave. We therefore decided to

The literacy and numeracy assessments measure core competencies that children should have achieved after two years of schooling according to the national curriculum. Literacy tests in both English and Swahili assess the following four competencies in order of rising difficulty: (1) recognition of letters, (2) recognition of words, (3) reading a paragraph, and (4) reading a short story. Numeracy tests measure the following six competencies in order of rising difficulty: (1) counting, (2) recognition of numbers, (3) rank ordering of numbers, (4) addition, (5) subtraction, and (6) multiplication. A student's score on each test equals the highest competency level achieved, with a zero indicating that she did not even master the simplest skill assessed. Previous analyses of Uwezo data have shown that a large proportion of students even in higher grades does not master these second-grade competencies (Jones et al., 2014; Uwezo, 2015).

3.2. Variable definitions

The key explanatory variable in our regressions is an indicator for whether a child has attended preschool or not. This variable is based on retrospectively reported information, which has the major advantage that we can estimate longer-term impacts by relating it to current outcomes. A potential concern with retrospective data is that it may be contaminated by recall error; specifically, if such recall error systematically depends on preschool attendance, this could lead to bias in our estimates (Garces, Thomas, and Currie, 2002). We therefore show in a robustness check that systematic recall error is unlikely to drive our results. For a subsample of children, we also observe for how many years they attended preschool. Later on, we use this information to estimate effects at the intensive margin.

Our empirical analysis focuses on two main outcomes: the highest school grade completed and a cognitive test score. We observe the highest grade completed both for children who are currently enrolled in school and for those who dropped out, with children who are still in preschool coded as having zero grades completed. Because all our regressions include age dummies, this variable is essentially a measure of school progression. We construct the cognitive test score by first standardizing the literacy (both

exclude Uganda from the analysis.

English and Swahili) and numeracy scores by country, Uwezo survey wave, and age to have mean zero and standard deviation one. In a second step, we then average these standardized scores for each student and standardize the resulting composite cognitive score again.

The control variables include a variety of socio-demographic characteristics, such as age and gender, mother's education, and an index of household wealth. Moreover, we construct two proxies for early-life economic conditions at the district level from external data sources. The first proxy is the log of average night light density, which has been shown to be a good measure of economic activity (Henderson, Storeygard, and Weil, 2012). The second proxy consists of two separate dummies for positive and negative rainfall shocks, defined as rainfall above the 80th percentile and below the 20th percentile of the long-term district mean, respectively. Rainfall shocks have been used widely as a measure of income shocks in rural economies; see Shah and Steinberg (2017) for a recent example. In order to allow for differential impacts of economic conditions at different ages, we compute our two proxies separately at each age before school entry (ages 0-5 in Kenya and ages 0-6 in Tanzania) for each child. In the Data Appendix, we provide many more details on the construction of these and all other variables used in the empirical analysis.

3.3. Sample selection and descriptive statistics

We use data from all available waves of the Uwezo surveys with information on preschool attendance; these are the 2013 and 2014 waves in Kenya and the four waves conducted between 2011 and 2014 in Tanzania. We restrict our attention to children aged 7/8 and above in Kenya/Tanzania because some younger children were still of preschool age at the time of the survey. In order to ensure that we focus on comparable siblings in our within-household analysis, we also drop from the sample any children who report never to have enrolled in preschool or school.[6] Our final sample comprises more than 500,000 children with information on preschool attendance and the two main outcomes described above. Table 1 reports

[6]Children of school age who have never enrolled in preschool or school include, for example, children with disabilities. Not surprisingly, if we include these children in our sample, the estimated returns to preschool are substantially higher.

descriptive statistics for this sample separately for Kenya and Tanzania.

4. Empirical strategy

4.1. Addressing selection bias

The main challenge in identifying the causal effects of preschool attendance on later outcomes is that selection into pre-primary education is likely non-random. For example, more educated parents may have a stronger preference for preschool education while also fostering their children's learning in other ways. In this case, a simple regression that does not control for this selection would yield estimates that are biased upward. To address this challenge, we follow a strand of previous literature (Currie and Thomas, 1995; Garces, Thomas, and Currie, 2002; Berlinski, Galiani, and Manacorda, 2008; Deming, 2009) and estimate models with household fixed effects, thus holding constant all determinants of preschool attendance and outcomes that do not vary between siblings:

$$Y_{ij} = \alpha + \beta PRE_{ij} \times \mathbf{AGE}_{ij} + \gamma \mathbf{AGE}_{ij} + \delta \mathbf{X}_{ij} + \eta_j + \varepsilon_{ij}. \quad (1)$$

Here, i indexes individuals and j indexes households, Y_{ij} is the highest grade completed or cognitive test score, PRE_{ij} is the indicator for whether the child has attended preschool, and $\mathbf{AGE}_{ij}$ is a series of age dummies. $\mathbf{X}_{ij}$ is a vector of controls that include cohort effects and their interaction with age, dummies for birth order and their interaction with gender, and all the socio-demographic characteristics and early-life economic conditions shown in Table 1. By interacting PRE_{ij} with $\mathbf{AGE}_{ij}$, we allow the effects of preschool attendance to vary by age; in practice, in order to avoid cluttering the tables, we report estimates from specifications in which PRE_{ij} is interacted with indicators for three age groups instead.

The regression in equation 1 identifies the causal effect of preschool attendance under the assumption that *among siblings*, selection into preschool is uncorrelated with any other determinants of the outcome. While comparatively weak, this assumption might be violated for several reasons, two of which are particularly salient. First, given that pre-primary education is costly, household income shocks around preschool age may be

driving siblings' differential enrollment. Because such income shocks influence longer-term educational success also in other ways (e.g. Shah and Steinberg, 2017), this might introduce bias into our estimates. We address this concern by including detailed controls for early-life economic conditions in our regressions. Second, siblings who attend preschool may differentially benefit also from further, unobserved educational inputs, for example due to parental favoritism. To mitigate this concern, we show in a robustness check that preschool attendance is not predictive of other costly investments such as private tutoring.

4.2. Sources of between-sibling variation in preschool attendance

Which factors drive the remaining between-sibling variation in preschool attendance after controlling for income shocks and parental favoritism? We investigate this question in Table 2, which reports results of regressions of the indicator for preschool attendance on the control variables. Columns 1 and 3 show estimates from a specification without household fixed effects for Kenya and Tanzania, respectively. In both countries, children of educated mothers and from richer households are more likely to have attended preschool, which points to the importance of controlling for between-family differences in our main specifications. Columns 2 and 4 show that once household fixed effects are included in the regressions, most of the variables that still vary between siblings are no longer predictive of preschool attendance, including the proxies for early-life economic conditions.[7]

The lower part of Table 2 reports the coefficients on the cohort effects. There is a marked and nearly monotonic trend in both countries, with cohorts born in later years being significantly more likely to have attended preschool. This trend is particularly strong in Tanzania, where attendance rates at baseline were much lower (see Section 2), and it survives even when household fixed effects are included in the regressions. These estimates sug-

[7]To avoid cluttering, rather than separate dummies for early-life economic conditions at each age, specifications in Table 2 simply include the number of positive and negative rainfall shocks and the average log night lights before school entry. Results are qualitatively similar if we include the full set of dummies instead; in particular, early-life economic conditions appear to be largely orthogonal to preschool attendance. Notably, this is not due to poor measurement, as these variables are highly predictive of our two main outcomes (results available upon request).

gest that the expansion of pre-primary education during our study period led to differences in preschool availability between siblings, which in turn are responsible for differences in attendance. As long as these changes in availability are unrelated to changes in other determinants of educational outcomes, this implies that we identify the true causal effects of preschool attendance in the analysis below.[8]

5. Results

5.1. Preschool attendance and school progression

Table 3 reports estimates of the effect of preschool attendance on highest grade completed for three age groups: up to 9 years old, 10-12 years old, and 13-16 years old. Panel A presents the results for Kenya, while panel B presents the results for Tanzania. Regressions in this and all other tables in the paper are weighted using the sampling weights provided with the Uwezo data, and standard errors are clustered at the district level.

Column 1 shows estimates of a parsimonious specification which only controls for age, cohort, and district fixed effects. Due to the frequent late enrollment in pre-primary and subsequently primary education, children who attend preschool initially have fewer grades completed than their peers who directly go to primary school. However, these children also progress through grades at a faster pace and eventually overtake their peers: at ages 13-16, children who attended preschool have completed 0.19 and 0.32 more grades in Kenya and Tanzania, respectively.

As discussed above, the results in column 1 are unlikely to reflect the causal impact of preschool attendance due to selection bias. As an intermediate step to mitigate this problem, we successively add controls for socio-demographic characteristics and early-life economic conditions to the regressions. Consistent with the idea of positive selection into pre-primary education, controlling for socio-demographic characteristics reduces the coefficients compared to column 1; in particular, the initial disadvantage from preschool attendance is now greater and the long-term benefits are

[8]Ideally, we would like to further investigate this hypothesis using data on preschool openings by district and year. Unfortunately, however, such data do not appear to exist.

smaller (column 2). Instead, in line with the conclusion above that early-life economic conditions are largely orthogonal to preschool attendance, controlling for them changes the coefficients only slightly (column 3).

Column 4 presents estimates from our preferred specification, which includes household fixed effects. Compared to the previous columns, the long-term benefits of preschool attendance are now further reduced, although still economically and statistically significant: in both countries, children who went to preschool accumulate about one and a half additional months (0.12 grades) of schooling by ages 13-16.[9] Panel A of Figure 1 plots estimates from regressions in which the effect of preschool attendance is allowed to differ at each age rather than at age groups. Beyond confirming the results in column 4 of Table 3, the plots show that the impact of preschool on grades completed rises almost monotonically with age.

An interesting question is whether children who attended preschool catch up with their peers because they skip more or repeat fewer grades while in school, or because they are less likely to drop out of school (recall that we measure the highest grade completed also for children who are no longer enrolled). We investigate this issue in column 1 of Appendix Table 1, which presents estimates from a regression of an indicator for having dropped out on preschool attendance. The results indicate that children who enrolled in pre-primary education are indeed less likely to drop out, especially at higher ages. Thus, lower dropout is partly underlying the gains in the number of grades completed shown in Table 3.[10]

[9]There are at least four explanations why the estimated long-term benefits of preschool attendance in column 4 are smaller than in the previous columns. First, unobserved household characteristics might bias the coefficients in the previous columns upward. Second, attenuation bias due to measurement error in attendance is aggravated in the between-sibling specification. Third, the inclusion of household fixed effects nets out any positive sibling spillovers. Fourth, the effects in column 4 are identified only from households with "within" variation, which might differ from those in the full sample. Investigating this last possibility, we found that households with both attending and non-attending children were larger, poorer, and more likely to be located in a rural area. However, when we restricted the sample to these households only, the inclusion of household fixed effects similarly led to a decline in the estimated long-term benefits of preschool. Later on, we also show that impacts do not differ between poorer and richer households. Together, this suggests that identification based on a different sample is not driving the change in coefficients between columns 3 and 4 of Table 3.

[10]Unfortunately, the Uwezo data do not contain information on school starting age and grade repetition, which prevents us from fully disentangling the mechanisms behind

5.2. Preschool attendance and cognitive test scores

Table 4 reports estimates of the effect of preschool attendance on cognitive test scores. Column 1 again shows results from a specification with only basic controls. In Kenya, children who attend preschool have a small advantage over their peers during the early ages, and this advantage grows to a sizable 0.1 SD for the two later age groups. In contrast, in Tanzania, children with pre-primary education outperform their peers by 0.26 SD already early on, but this difference decreases to 0.22 SD for the oldest age group of 13-16 year-old children.

In columns 2-4, we successively add our control variables and household fixed effects. In Kenya, the estimates change only little with the inclusion of these variables. In contrast, the coefficients in Tanzania are substantially reduced, with the impact for the oldest age group now estimated at 0.08 SD. Panel B of Figure 1 plots estimates of specifications in which the effect of preschool attendance is allowed to differ at each age. The plots confirm the patterns of a fade-in and later stabilization of the preschool impact in Kenya, and the initial positive effect that fades out slightly in Tanzania. Finally, columns 2-4 of Appendix Table 1 report results from regressions of the individual numeracy, English, and Swahili scores which are very similar to the ones for the summary cognitive test score.

Taken together, the results in Tables 3 and 4 show an interesting pattern of how preschool attendance impacts learning outcomes across age groups. In both Kenya and Tanzania, students who enroll in pre-primary education tend to start primary school later than their peers, but they eventually catch up and at later ages even outperform them in terms of grades completed. This higher attainment is accompanied by substantial gains in cognitive skills, which materialize early on, especially in Tanzania. Overall, there are thus important long-term benefits from preschool attendance for children's educational outcomes.

the catch-up of children who attended preschool. In Appendix Table 1, note that for the younger two age groups, dropout rates are very low and the estimated coefficients from the linear probability model sometimes exceed the outcome mean. Addressing this issue, we confirmed that probit models and a simple comparison of means also suggest that children who attended preschool are less likely to drop out of school.

5.3. Heterogeneity

We now explore the heterogeneity of the preschool impacts along several dimensions. First, the indicator for having attended preschool used in the regressions so far does not distinguish between different lengths of attendance. Instead, Figure 2 plots estimates from specifications in which this indicator is replaced by separate dummies for having attended preschool for 1, 2, or 3 years. Panel A shows that in both countries, children who attend preschool for longer are initially further behind their peers in terms of grades completed, likely because they start school later. However, these children also catch up faster and at ages 13-16, there is no longer any significant difference between children with different lengths of attendance.[11] Panel B shows that in terms of test scores, there appears to be comparatively little heterogeneity by length of attendance; note, however, that the confidence intervals in these regressions are substantially larger.

Second, we examine whether preschool impacts differ by the characteristics of children or households. Figure 3 plots estimates from regressions in which the indicator for preschool attendance is interacted with a gender dummy. While the results suggest that girls profit slightly more from preprimary education, this difference is never statistically significant. Figure 4 plots estimates from specifications that allow the coefficient on attendance to vary with household wealth, which reveal no consistent differences in impacts between richer and poorer households. Finally, in unreported regressions, we also examined whether preschool impacts differ by mother's education and rural versus urban location, but similarly found no consistent differences between groups (results are available upon request).

5.4. Robustness

In this subsection, we address several potential concerns regarding the interpretation of our results. First, we argue above that the between-sibling variation in preschool attendance is likely due to changes in availability, which came about because of the expansion of the pre-primary sector during

[11]The effect sizes for Tanzania in Figure 2 differ somewhat from those reported in Table 3. The reason is that the sample now excludes the 2011 and 2012 waves of the Uwezo surveys, which do not contain information on length of preschool attendance.

our study period. One might worry, however, that this variation instead reflects child-specific investments which are correlated with unobservables. If this is indeed the case, one can reasonably expect that families differentially invest in children who are sent to preschool also in other ways. We test this hypothesis by examining whether children who attended preschool are more likely to benefit from two other costly investments observed in our data: private after-school tutoring and enrollment in private school.[12] Columns 1 and 2 of Table 5 show that in regressions with indicators for receiving these investments as outcomes, the coefficients on preschool attendance are always close to zero and very precisely estimated. This suggests that differential investments based on unobservable child characteristics are not driving our results.

Second, another potential worry is that our results are driven by siblings who are very different in age, and who thus grew up under very distinct circumstances. To alleviate this concern, in columns 3 and 4 of Table 5, we show that estimates are very similar when the sample is restricted to families in which all children are born at most 5 years apart. Similarly, columns 5 and 6 of Table 5 report results from regressions in which the sample is restricted to sibling pairs only. While this restriction reduces the number of observations substantially, the estimates are largely unchanged.

Finally, we address the concern that our results are affected by recall error in the retrospectively reported preschool variable. Of course, if such recall error exists but is unrelated to preschool attendance, this will simply bias our estimates towards zero. Thus, it introduces bias only if it systematically varies between siblings who did and did not attend preschool. We tackle this issue by taking advantage of the repeated cross-sectional nature of our data, which lets us follow cohorts over time. Table 6 reports the fraction of children in each cohort reporting to have attended preschool separately for each wave in our data.[13] For the vast majority of cohorts,

[12]See Wamalwa and Burns (2017) for an analysis of private school effectiveness in Kenya using the Uwezo survey data.

[13]For this exercise, we focus on a comparable sample of districts that were visited in all waves of the Uwezo surveys. We disregard the 2014 wave in Tanzania because only a small subsample of districts were included in that year's survey. In the raw data, we observe level shifts in preschool attendance rates *for all cohorts* between some of the waves, likely because the question asking about preschool attendance changed.

these fractions do not change substantially between waves, suggesting that systematic recall error does not bias our estimates.

6. Conclusion

Most children in Sub-Saharan Africa enroll in school nowadays, but they are learning remarkably little there. One possible reason is that they enter school unprepared, which makes preschool programs that aim to get children ready for school a promising way to improve learning outcomes. While such pre-primary education is increasingly common in the region, very little is known about its effectiveness.

In this paper, we provide some of the first evidence of preschool impacts on learning outcomes in Sub-Saharan African, and in developing countries more generally. We use data from large-scale surveys of children's educational attainment and cognitive skills in Kenya and Tanzania, which also collect retrospective information on preschool attendance. Our analysis compares the highest grade completed and cognitive skills, given age, of siblings who did and did not attend preschool. This strategy allows us to control for any determinants of pre-primary enrollment and outcomes that do not vary within households. We provide evidence that the leftover between-sibling variation in attendance is due to changes in availability, which came about because of a large expansion of pre-primary education during our study period.

The results show that preschool attendance leads to important long-term benefits for children: at ages 13-16, they have completed about one and a half more months of schooling, and they outperform their peers on cognitive tests by 0.12 SD and 0.08 SD in Kenya and Tanzania, respectively. These gains, especially in terms of grades completed, materialize relatively late because children who attend preschool tend to enter primary school late and thus fall behind early on. While we can only speculate on the mechanisms behind these findings, the most likely explanation is that children learn basic cognitive or behavioral skills in preschool, which help them le-

In Table 6, we therefore report regression-adjusted attendance rates after taking out wave fixed effects. Note that level shifts in preschool attendance do not influence our within-household results, which use variation within survey waves.

arn more effectively in primary school. Overall, our analysis shows that increasing access to pre-primary education can be an effective instrument to improve learning outcomes in Sub-Saharan Africa.

References

Behrman, J.R., Y. Cheng, and P.E. Todd. 2004. "Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach." *Review of Economics and Statistics* 86:108–132.

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

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

Bidwell, K., K. Parry, and L. Watine. 2013. *Exploring Early Education Programs in Peri-urban Settings in Africa: Nairobi Report*. Report, Innovations for Poverty Action.

Bietenbeck, J., M. Piopiunik, and S. Wiederhold. 2017. "Africa's Skill Tragedy: Does Teachers' Lack of Knowledge Lead to Low Student Performance?" *Journal of Human Resources* forthcoming.

Bouguen, A., D. Filmer, K. Macours, and S. Naudeau. 2017. "Preschool and Parental Response in a Second Best World: Evidence from a School Construction Experiment." *Journal of Human Resources* forthcoming.

Cunha, F., and J. Heckman. 2007. "The Technology of Skill Formation." *American Economic Review* 97:31–47.

Currie, J., and D. Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85:341–64.

Deming, D. 2009. "Early childhood intervention and life-cycle skill development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1:111–134.

Edwards Jr., D.B., S.J. Klees, and J. Wildish. 2015. "Dynamics of Low-Fee Private Schools in Kenya: Governmental Legitimation, School-Community Dependence, and Resource Uncertainty." *Teachers College Record* forthcoming.

Garces, E., D. Thomas, and J. Currie. 2002. "Longer-Term Effects of Head Start." *American Economic Review* 92:999–1012.

Henderson, J.V., A. Storeygard, and D. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102:944–1028.

Heyneman, S.P., and J.M. Stern. 2014. "Low cost private schools for the poor: What public policy is appropriate?" *International Journal of Educational Development* 35:3 – 15.

Jones, S., Y. Schipper, S. Ruto, and R. Rajani. 2014. "Can your child read and count? Measuring learning outcomes in East Africa." *Journal of African Economies* 23:643–672.

Martinez, S., S. Naudeau, and V. Pereira. 2013. "The Promise of Preschool in Africa: A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique." Working paper.

Mligo, I.R. 2016. "Teachers' perceptions and concerns about the implementation of the 2005 preschool curriculum in Tanzania." *Early Years* 36:353–367.

Piper, B., S.S. Zuilkowski, and A. Mugenda. 2014. "Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative." *International Journal of Educational Development* 37:11–21.

Shah, M., and B.M. Steinberg. 2017. "Drought of Opportunities: Contemporaneous and Long-Term Impacts of Rainfall Shocks on Human Capital." *Journal of Political Economy* 125:527–561.

Tooley, J., P. Dixon, and J. Stanfield. 2008. "Impact of Free Primary Education in Kenya: A Case Study of Private Schools in Kibera." *Educational Management Administration & Leadership* 36:449–469.

UNESCO. 2012. *Global education digest 2012. Opportunities lost: The impact of grade repetition and early school leaving*. Paris: UNESCO.

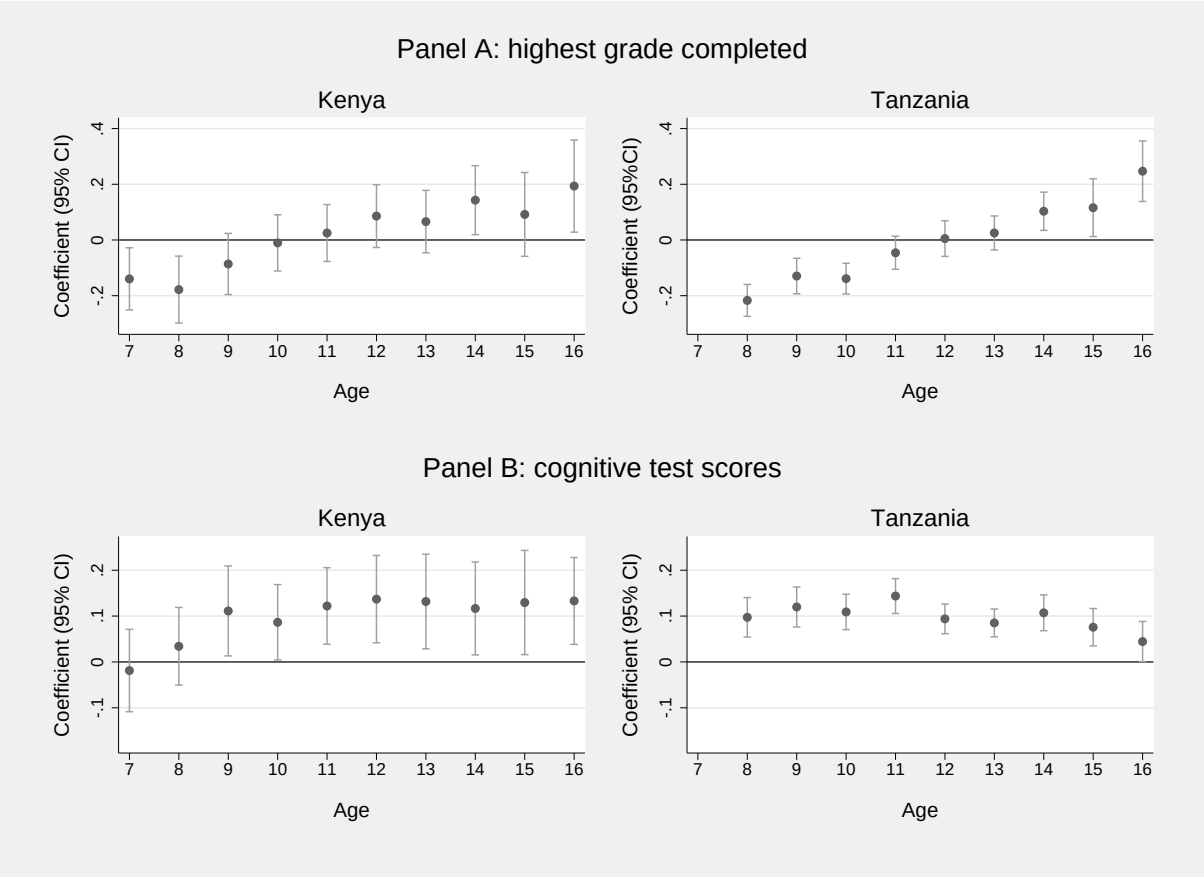
Uwezo. 2015. *Are Our Children Learning? Literacy and Numeracy Across East Africa 2014*. Nairobi: Twaweza East Africa.

Wamalwa, F.M., and J. Burns. 2017. "Private Schools and Student Learning Achievements in Kenya." SALDRU Working Paper No. 202.

World Bank. 2012. *SABER early childhood development country report: Tanzania 2012*. Washington, DC: World Bank.

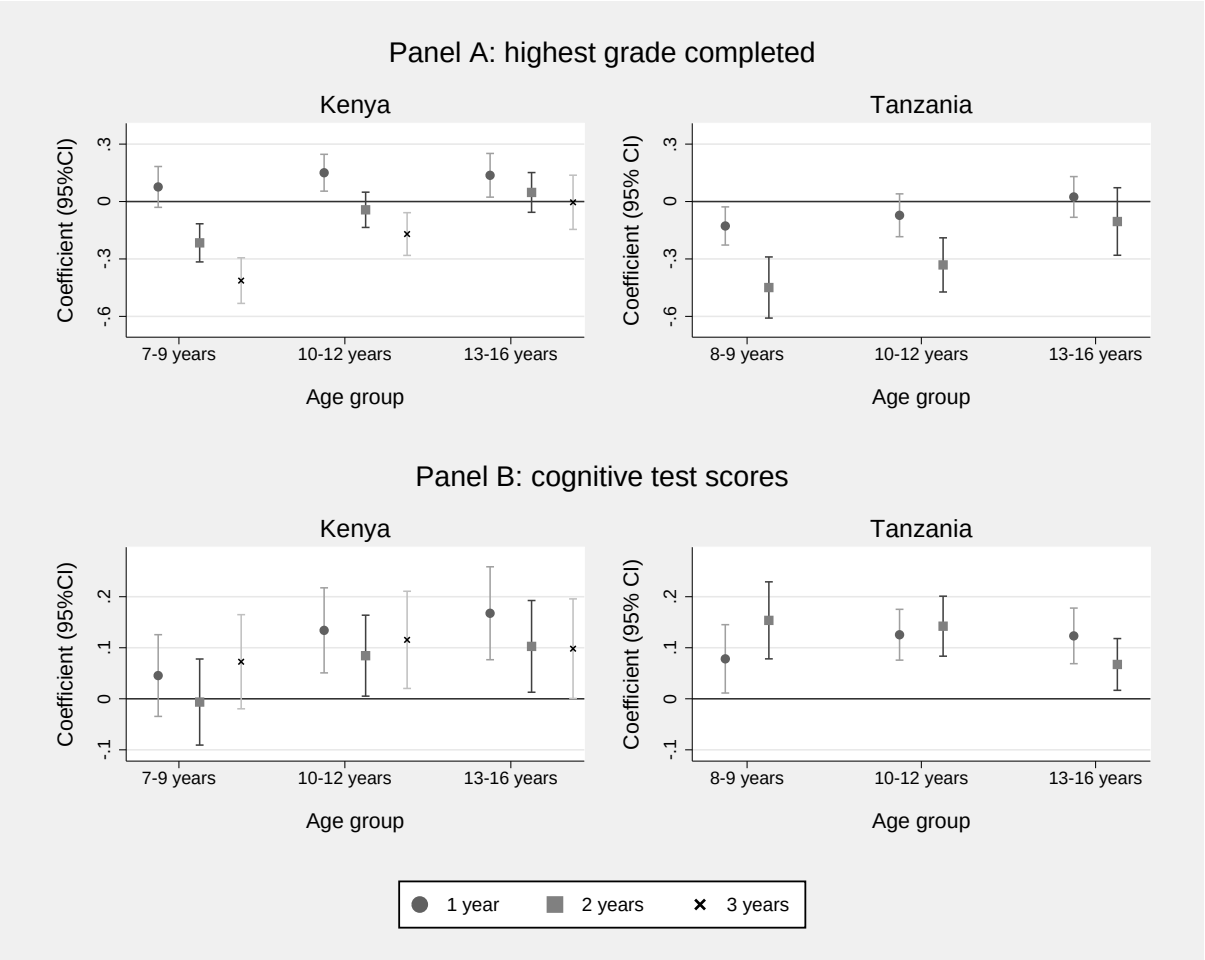
—. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. Washington, DC: World Bank.

Figure 1
Age-specific impacts of preschool attendance



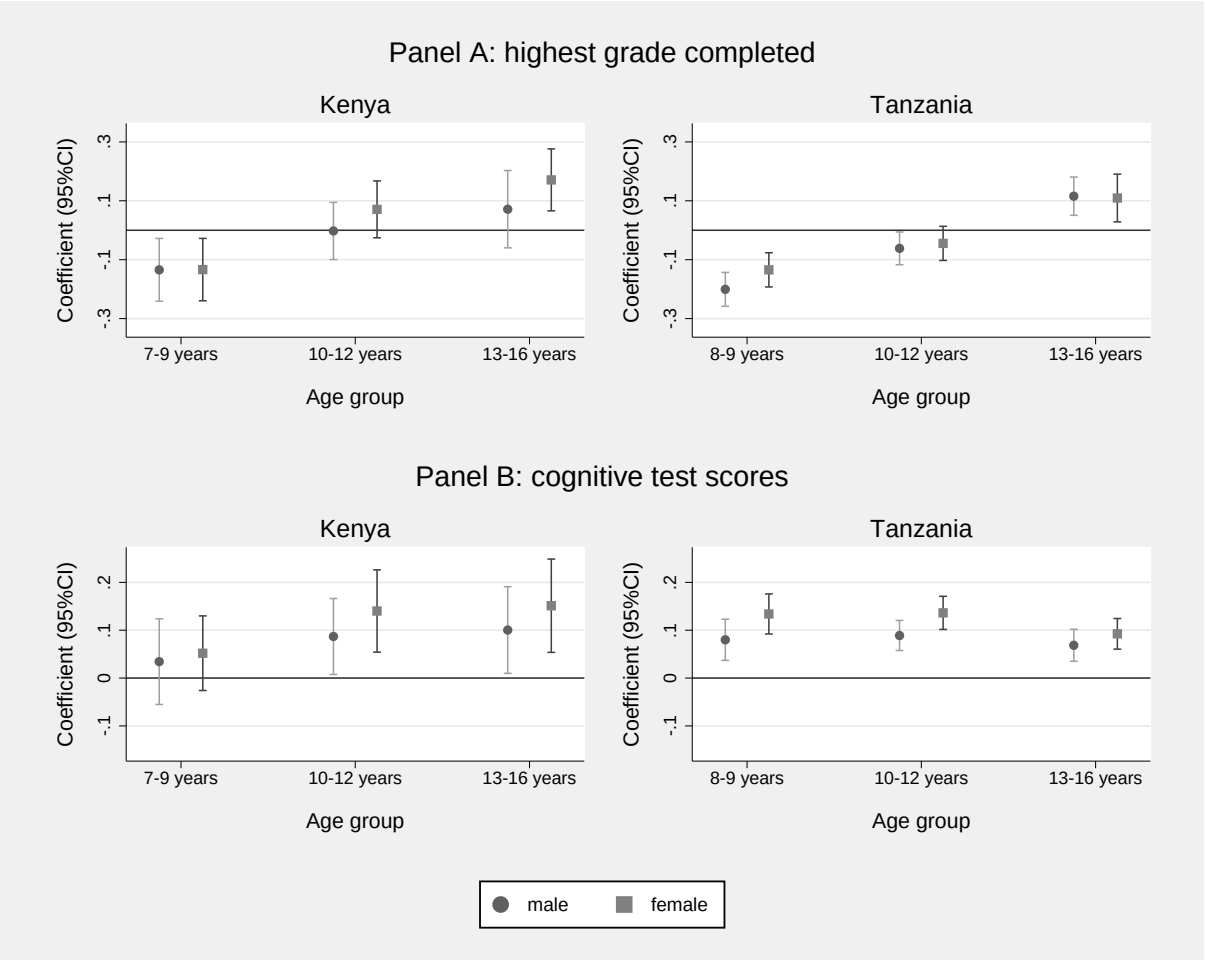
Notes: The figure shows estimates of age-specific impacts of preschool attendance on the highest grade completed (panel A) and the cognitive test score (panel B). Panel A plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 3 in which the indicator for preschool attendance is interacted with individual age dummies rather than age-group dummies. Panel B plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 4 in which the indicator for preschool attendance is interacted with individual age dummies rather than age-group dummies.

Figure 2
Heterogeneity by years of preschool attended



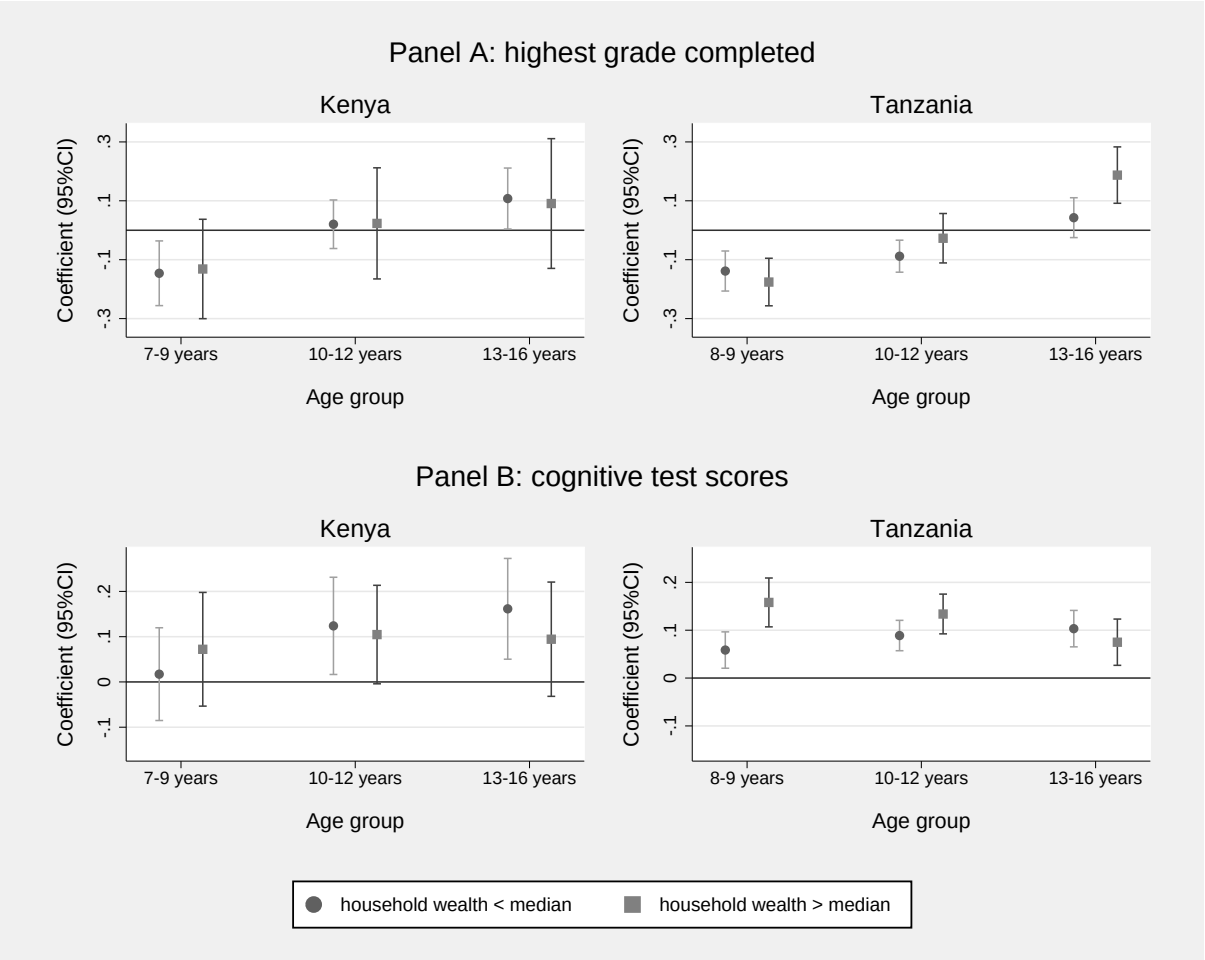
Notes: The figure shows estimates from regressions which probe for heterogeneity of the impact of preschool attendance by length of attendance. Panel A plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 3 in which the indicator for preschool attendance is replaced with separate indicators for having attended 1, 2, and 3 years of preschool. Panel B plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 4 in which the indicator for preschool attendance is replaced with separate indicators for having attended 1, 2, and 3 years of preschool. Years of attendance are only observed in the 2013 and 2014 waves of Uwezo, meaning that the 2011 and 2012 waves for Tanzania are dropped from the sample for this analysis.

Figure 3
Heterogeneity by gender



Notes: The figure shows estimates from regressions which probe for heterogeneity of the impact of preschool attendance by child gender. Panel A plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 3 in which the indicator for preschool attendance is interacted with a dummy for female. Panel B plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 4 in which the indicator for preschool attendance is interacted with a dummy for female. Plots show the overall effects for boys and girls.

Figure 4
Heterogeneity by household wealth



Notes: The figure shows estimates from regressions which probe for heterogeneity of the impact of preschool attendance by household wealth. Panel A plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 3 in which the indicator for preschool attendance is interacted with a dummy household wealth above the country median. Panel B plots coefficient estimates and 95% confidence intervals from variations of the specifications in column 4 of Table 4 in which the indicator for preschool attendance is interacted with a dummy household wealth above the country median. Plots show the overall effects for below-median-wealth and above-median-wealth households.

Table 1
Summary statistics

	Kenya	Tanzania
Socio-demographic characteristics		
Age	11.08	11.72
	(2.77)	(2.46)
Female	0.49	0.50
	(0.50)	(0.50)
Mother's education		
None	0.17	0.19
	(0.37)	(0.39)
Some primary or more	0.83	0.81
	(0.37)	(0.39)
No. of children in household	3.09	2.47
	(1.55)	(1.26)
Household wealth index	0.00	0.00
	(1.00)	(1.00)
Rural location	0.67	0.78
	(0.47)	(0.41)
Early-life economic conditions		
No. of negative rainfall shocks	1.44	1.87
	(0.85)	(0.87)
No. of positive rainfall shocks	1.10	1.27
	(0.79)	(0.72)
Log night light density	-1.44	-2.39
	(2.32)	(2.29)
Preschool attendance		
Attended preschool	0.85	0.62
	(0.36)	(0.48)
Years of preschool	1.79	0.67
	(1.08)	(0.75)
Outcomes		
Highest grade completed	3.67	3.45
	(2.49)	(2.28)
Cognitive test score	0.00	0.00
	(1.00)	(1.00)
Observations	223,339	293,757
% with within variation	3.37	10.61

Notes: The table reports means and standard deviations (in parentheses) of key variables separately for children in Kenya and Tanzania. Variables measuring early-life economic conditions are computed across ages 0-5 in Kenya and ages 0-6 in Tanzania. Number of negative (positive) rainfall shocks counts the number of years in which rainfall was below the 20th percentile (above the 80th percentile) of the long-term district mean. Log night light density is computed as the district average across ages. Years of preschool attended is observed only for part of the sample. The final row reports the fraction of children living in households in which at least one child went to preschool and at least one child did not. Further details on the construction of all variables used in the empirical analysis are provided in the Data Appendix.

Table 2
Predicting preschool attendance

	Kenya		Tanzania	
	(1)	(2)	(3)	(4)
Female	0.002	-0.001	0.006**	0.003
	(0.002)	(0.002)	(0.003)	(0.003)
Firstborn	0.005	0.001	-0.004	0.002
	(0.003)	(0.002)	(0.003)	(0.004)
Mother $\geq$ some primary	0.016**		0.097***	
	(0.007)		(0.006)	
No. of children in household	-0.002		-0.012***	
	(0.002)		(0.002)	
Household wealth index	0.006**		0.052***	
	(0.003)		(0.004)	
Rural location	-0.000*		-0.016	
	(0.000)		(0.022)	
No. of negative rainfall shocks	-0.003**	-0.001	-0.007*	-0.000
	(0.001)	(0.001)	(0.004)	(0.004)
No. of positive rainfall shocks	-0.004	-0.000	0.006	0.007*
	(0.002)	(0.001)	(0.004)	(0.004)
Log night light density	0.009	0.002	0.016	0.011
	(0.006)	(0.005)	(0.016)	(0.012)
Cohort				
1996			0.023***	0.021
			(0.008)	(0.014)
1997			0.030***	0.035***
			(0.009)	(0.011)
1998	0.005	0.002	0.047***	0.050***
	(0.008)	(0.004)	(0.009)	(0.013)
1999	0.012*	0.004	0.062***	0.070***
	(0.007)	(0.004)	(0.010)	(0.014)
(...)				
2004	0.023**	0.007	0.113***	0.110***
	(0.009)	(0.005)	(0.012)	(0.017)
2005	0.033***	0.008	0.099***	0.099***
	(0.009)	(0.006)	(0.012)	(0.018)
2006	0.040***	0.015**	0.147***	0.149***
	(0.007)	(0.006)	(0.017)	(0.023)
2007	0.032***	0.014		
	(0.009)	(0.009)		
District fixed effects	Yes	No	Yes	No
Household fixed effects	No	Yes	No	Yes
Observations	223,339	223,339	293,757	293,757

Notes: The table reports estimates from regressions of an indicator for preschool attendance on the variables listed in rows and Uwezo wave dummies. Standard errors in parentheses are clustered at the district level. * p<0.10, ** p<0.05, *** p<0.01.

Table 3
Preschool attendance and highest grade completed

	(1)	(2)	(3)	(4)
	Panel A: Kenya			
Attended preschool				
7-9 years old	-0.158***	-0.174***	-0.164***	-0.135***
	(0.041)	(0.043)	(0.042)	(0.049)
10-12 years old	0.036	0.014	0.014	0.035
	(0.023)	(0.022)	(0.022)	(0.046)
13-16 years old	0.185***	0.170***	0.161***	0.120**
	(0.035)	(0.035)	(0.033)	(0.055)
Observations	218,728	218,728	218,728	218,728
	Panel B: Tanzania			
Attended preschool				
8-9 years old	-0.032*	-0.087***	-0.075***	-0.169***
	(0.017)	(0.017)	(0.017)	(0.026)
10-12 years old	0.093***	0.039**	0.042**	-0.053**
	(0.019)	(0.018)	(0.018)	(0.024)
13-16 years old	0.315***	0.261***	0.245***	0.114***
	(0.026)	(0.026)	(0.024)	(0.031)
Observations	284,396	284,396	284,396	284,396
	Controls included in panels A and B			
District fixed effects	Yes	Yes	Yes	No
Socio-dem. characteristics	No	Yes	Yes	Yes
Early-life econ. conditions	No	No	Yes	Yes
Household fixed effects	No	No	No	Yes

Notes: The table reports estimates from regressions of the highest grade completed on an indicator for preschool attendance and control variables as indicated in the lower panel. In the regressions, the indicator for preschool attendance is interacted with three age-group dummies (7/8-9, 10-12, and 13-16 years), and the table reports the estimated effect of preschool attendance separately for each group. All specifications include dummies for age and cohort, their interactions, and Uwezo wave dummies. Socio-demographic characteristics include the variables reported in Table 1 as well as dummies for birth order and their interactions with gender. Controls for early-life economic conditions are interacted with an indicator for rural location. Standard errors in parentheses are clustered at the district level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 4

Preschool attendance and cognitive test scores

	(1)	(2)	(3)	(4)
	Panel A: Kenya			
Attended preschool				
7-9 years old	0.036*	0.015	0.012	0.042
	(0.022)	(0.022)	(0.022)	(0.040)
10-12 years old	0.097***	0.078***	0.077***	0.114***
	(0.022)	(0.021)	(0.021)	(0.040)
13-16 years old	0.104***	0.095***	0.098***	0.124***
	(0.021)	(0.021)	(0.021)	(0.044)
Observations	218,134	218,134	218,134	218,134
	Panel B: Tanzania			
Attended preschool				
8-9 years old	0.256***	0.204***	0.195***	0.107***
	(0.017)	(0.017)	(0.015)	(0.018)
10-12 years old	0.263***	0.212***	0.209***	0.113***
	(0.014)	(0.013)	(0.013)	(0.014)
13-16 years old	0.217***	0.163***	0.170***	0.080***
	(0.016)	(0.015)	(0.014)	(0.014)
Observations	288,084	288,084	288,084	288,084
	Controls included in panels A and B			
District fixed effects	Yes	Yes	Yes	No
Socio-dem. characteristics	No	Yes	Yes	Yes
Early-life econ. conditions	No	No	Yes	Yes
Mother fixed effects	No	No	No	Yes

Notes: The table reports estimates from regressions of the cognitive test score on an indicator for preschool attendance and control variables as indicated in the lower panel. In the regressions, the indicator for preschool attendance is interacted with three age-group dummies (7/8-9, 10-12, and 13-16 years), and the table reports the estimated effect of preschool attendance separately for each group. All specifications include dummies for age and cohort, their interactions, and Uwezo wave dummies. Socio-demographic characteristics include the variables reported in Table 1 as well as dummies for birth order and their interactions with gender. Controls for early-life economic conditions are interacted with an indicator for rural location. Standard errors in parentheses are clustered at the district level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 5
Robustness

	Falsification tests		Born ≤5 years apart		Sibling pairs	
	Private tutoring (1)	Private school (2)	Highest grade (3)	Cognitive score (4)	Highest grade (5)	Cognitive score (6)
	Panel A: Kenya					
Attended preschool						
7-9 years old	-0.002	-0.003	-0.061	0.066	-0.019	0.010
	(0.011)	(0.010)	(0.067)	(0.052)	(0.066)	(0.060)
10-12 years old	0.004	-0.013	0.069	0.136***	0.140**	0.091
	(0.011)	(0.011)	(0.071)	(0.051)	(0.066)	(0.061)
13-16 years old	0.020	-0.015	0.096	0.122**	0.255***	0.086
	(0.017)	(0.011)	(0.079)	(0.054)	(0.066)	(0.067)
Observations	223,339	208,424	158,132	158,408	67,970	67,995
	Panel B: Tanzania					
Attended preschool						
8-9 years old	-0.001	-0.005	-0.159***	0.100***	-0.174***	0.122***
	(0.024)	(0.004)	(0.027)	(0.021)	(0.037)	(0.020)
10-12 years old	0.017	-0.006*	-0.059*	0.119***	-0.040	0.143***
	(0.024)	(0.003)	(0.031)	(0.016)	(0.035)	(0.019)
13-16 years old	0.019	-0.001	0.077**	0.077***	0.118**	0.094***
	(0.019)	(0.003)	(0.033)	(0.018)	(0.048)	(0.022)
Observations	25,346	264,810	235,967	238,852	101,317	102,485

Notes: Columns 1 and 2 report estimates from regressions in which the dependent variables are indicators for receiving private tutoring (only observed in 2014 in Tanzania) and for current enrollment in private school, respectively. The specifications are otherwise identical to the ones in column 4 of Tables 3 and 4. Means of the dependent variables for the youngest/middle/oldest age group: Kenya – private tutoring: 0.27/0.31/0.37; Tanzania – private tutoring: 0.22/0.25/0.25; Kenya – private school: 0.24/0.15/0.09; Tanzania – private school: 0.06/0.05/0.08. Columns 3 and 4 report variations from the regressions in column 4 of Table 3 and column 4 of Table 4, respectively, in which the sample is restricted to families in which all children are born at most 5 years apart. Columns 5 and 6 report variations from the regressions in column 4 of Table 3 and column 4 of Table 4, respectively, in which the sample is restricted to sibling pairs. Standard errors in parentheses are clustered at the district level. * p<0.10, ** p<0.05, *** p<0.01.

Table 6

Reported preschool attendance by cohort and wave

Wave:	2011	2012	2013	2014
		Panel A: Kenya		
Cohort				
1997			0.83	
1998			0.84	0.83
1999			0.84	0.84
2000			0.83	0.84
2001			0.84	0.85
2002			0.85	0.84
2003			0.84	0.85
2004			0.85	0.84
2005			0.85	0.85
2006			0.87	0.85
2007				0.85
		Panel B: Tanzania		
Cohort				
1995	0.57			
1996	0.59	0.58		
1997	0.59	0.58	0.60	
1998	0.61	0.60	0.61	
1999	0.63	0.61	0.62	
2000	0.65	0.64	0.64	
2001	0.66	0.65	0.63	
2002	0.69	0.66	0.65	
2003	0.69	0.67	0.64	
2004		0.68	0.65	
2005			0.64	

Notes: The table shows the fractions of children reporting to have attended preschool by country, cohort, and Uwezo survey wave. The sample is restricted to districts that were visited in all waves of the Uwezo survey. We disregard the 2014 wave in Tanzania because only a subsample of districts were sampled (45 districts versus more than 120 districts in the three previous waves). In the raw data, we observe level shifts in preschool attendance rates *for all cohorts* between some of the waves, likely because the question asking about preschool attendance changed. The table therefore shows regression-adjusted attendance rates after taking out wave fixed effects.

Appendix Table 1
Results for further outcomes

	Dropout	Numeracy score	English score	Swahili score
	(1)	(2)	(3)	(4)
	Panel A: Kenya			
Attended preschool				
7-9 years old	-0.011**	0.029	0.047	0.033
	(0.005)	(0.041)	(0.032)	(0.038)
10-12 years old	-0.013**	0.098**	0.096***	0.093**
	(0.005)	(0.038)	(0.034)	(0.041)
13-16 years old	-0.020***	0.109**	0.112***	0.082*
	(0.006)	(0.044)	(0.035)	(0.045)
Observations	223,339	214,070	215,528	213,187
	Panel B: Tanzania			
Attended preschool				
8-9 years old	-0.045***	0.126***	0.030*	0.120***
	(0.005)	(0.019)	(0.018)	(0.017)
10-12 years old	-0.051***	0.116***	0.052***	0.118***
	(0.005)	(0.014)	(0.015)	(0.013)
13-16 years old	-0.090***	0.062***	0.072***	0.056***
	(0.006)	(0.016)	(0.014)	(0.014)
Observations	293,757	284,358	283,618	282,994

Notes: The table reports estimates from regressions of an indicator for having dropped out of school (column 1) and subject-specific test scores (columns 2-4) on an indicator for preschool attendance. All specifications include household fixed effects and controls as indicated in column 4 of Tables 3 and 4. Means of the dropout variable for the youngest/middle/oldest age group: 0.003/0.005/0.025 (Kenya) and 0.020/0.031/0.111 (Tanzania). All subject-specific test scores are standardized to have mean 0 and standard deviation 1. Standard errors in parentheses are clustered at the district level. * p<0.10, ** p<0.05, *** p<0.01.

DATA APPENDIX TO:

Preschool Attendance, School Progression, and Cognitive Skills in East Africa

Jan Bietenbeck

Sanna Ericsson

Fredrick M. Wamalwa

1. The Uwezo surveys: sampling and test design

Uwezo, which means 'capability' in Kiswahili, is a non-governmental organization that aims to improve competencies in literacy and numeracy among school-aged children in East Africa. Since 2009, Uwezo has conducted annual assessments of the basic literacy and numeracy skills of children in Kenya, Tanzania, and Uganda. The assessments are administered as part of repeated cross-sectional household surveys, which also collect information on a variety of child and household characteristics and education outcomes. Households are selected in a two-stage sampling design: first, in each census district of each country, 30 enumeration areas (which typically correspond to one or several villages) are sampled with probability proportional to size; then, 20 households in each of these enumeration areas are randomly selected to participate in the survey.[1] The resulting sample is representative at both the national and the district level. Weights which reflect this sampling design and which implement a number of ex-post corrections are provided with the data; we use these weights throughout our analysis.[2]

In participating households, all children aged 6-16 (7-16 in Tanzania) are assessed on core literacy and numeracy competencies that should be achieved after two years of schooling according to the national curriculum. Two separate literacy tests in English and Swahili measure the following four competencies in order of rising difficulty: (1) recognition of letters, (2) recognition of words, (3) reading a paragraph, and (4) reading a short story. The numeracy test measures the following six competencies in order of rising difficulty: (1) counting, (2) recognition of numbers, (3) rank ordering

[1] A few districts were excluded in some rounds of the survey due to security concerns. In 2014, Tanzania selected households from a random subsample of districts only.

[2] Unfortunately, the weights included with the Tanzanian data over-emphasize the importance of observations in 2014. Specifically, as reported in the previous footnote, only a random subsample of districts was surveyed in 2014, and this wave correspondingly includes less than a third of the observations compared to any previous wave. Nevertheless, the weights in 2014 add up to about 125% of the weights in all previous waves. We attempt to correct for this irregularity by re-scaling the 2014 weights at the district level, using the relative importance of each district in the 2013 wave as a scaling factor. Our results are however robust to using the original weights, not using any weights at all, or dropping the 2014 wave for Tanzania altogether.

of numbers, (4) addition, (5) subtraction, and (6) multiplication.[3] For each assessment, there are several test booklets in order to prevent children within the same household from copying each other's answers. A child's score on each test equals the highest competency level achieved, with a zero indicating that she did not even master the simplest skill assessed.

2. Variable definitions

Household identifier. The data contain a household identifier, which we use to construct household-level variables such as number of children and wealth. Because polygamy is common in some communities, a few households contain children from different mothers. For each child, we observe his/her mother's age and education, which we use to construct a unique mother identifier. Our within-household specifications are based on this more conservative mother identifier rather than the household identifier, even though in practice this makes little difference.

Socio-demographic characteristics. We define a child's cohort as Uwezo survey wave minus age. Mother's education is recorded differently between countries and survey waves; we make this variable comparable by collapsing it into two categories: no education and at least some primary education. To construct the household wealth index, we follow Schady et al. (2015) and aggregate the following dwelling characteristics and assets using the first principal component: wall materials, source of lighting, tv, radio, computer (only Kenya), mobile phone, bicycle, motorbike, and motor vehicle. We compute this index separately for each country and normalize it to have mean zero and standard deviation one.

The rural indicator describes the location of the enumeration area. For Kenya, this variable is not included with the publicly available data, but we were able to obtain it directly from Uwezo. For Tanzania, the variable is included in the publicly available data for the 2011 and 2012 survey waves; as we were not able to obtain the variable for the 2013 and 2014 waves, it

[3] In Kenya, children who master multiplication are also assessed on their division skills. We ignore this seventh, higher competency here in order to ensure comparability of test scores across Kenya and Tanzania.

is missing for children observed in these years.[4]

Early-life economic conditions. We construct two proxies for district-level economic conditions using external satellite data on night lights and rainfall. For Kenya, district definitions in the Uwezo data are based on the 2009 census. For Tanzania, we create a crosswalk which maps districts in the Uwezo data to districts in the 2002 census. We use GIS census district boundary files from IPUMS International to compute summary statistics for our two proxies for each district and year.[5]

We obtain the night lights data from the Defense Meteorological Satellite Program's Operational Linescan System (DMSP-OLS).[6] The data provide yearly measurements of average light density at a fine geographical level, with light density ranging from 0 to 63. For a detailed description of these data, we refer to Henderson, Storeygard, and Weil (2012). Our rainfall measures are derived from the Climate Hazards group Infrared Precipitation with Stations (CHIRPS) data.[7] These data provide annual measures of precipitation since 1981; for details, see Funk et al. (2015).

From the satellite data, we construct a variable measuring average log night lights and indicators for positive and negative rainfall shocks at each age before school entry (ages 0-5 in Kenya and ages 0-6 in Tanzania). In line with recent literature (e.g. Shah and Steinberg, 2017), we define rainfall shocks as precipitation above the 80th percentile and below the 20th percentile of the long-term district mean.

Preschool attendance. Recent waves of the Uwezo survey ask respondents whether they ever attended preschool and whether they are currently

[4] As usual in survey data, there are some missing values also in other control variables. In order not to unnecessarily reduce sample size, we impute missing values at the sample mean and include separate dummies for missing values on each control variable in all of our regressions.

[5] The district boundary files for the Kenyan 2009 census and the Tanzanian 2002 census are available here: https://international.ipums.org/international/gis_yrspecific_2nd.shtml.

[6] We use the Average Visible, Stable Lights, and Cloud Free Coverages series, which is available here: https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.

[7] We use the CHIRPS-2.0 global annual yearly data series, which is available here: ftp://ftp.chg.ucsb.edu/pub/org/chg/products/CHIRPS-2.0/global_annual/tifs/.

still enrolled in preschool. From the answers to these two question, we construct our key explanatory variable as an indicator which takes value 1 if a child ever attended preschool and 0 otherwise. In the 2013 and 2014 waves, we moreover have information on length of attendance in years. As a few respondents indicate lengths of attendance far beyond the usual, we winsorize this variable at 3 years in Kenya and 2 years in Tanzania (i.e. at the maximum "normal" length according to the national education system).

In Section 2 of the paper, we report preschool enrollment rates for different cohorts in our data. We focus on cohorts aged at least 10 years in order to account for the frequent late enrollment in preschool. The figures reported in that Section are based on the raw data and therefore differ somewhat from those reported in Table 6, which are based on our analysis sample (the figures in Table 6 moreover account for level shifts in reported enrollment across waves, see Section 5 for details).

Outcome variables. Our first main outcome is the highest grade of school completed. Children who are currently enrolled in preschool are coded as having zero grades completed. Children who are currently in school report the grade they are attending; for them, the highest grade completed equals the current grade minus one. Children who have dropped out of school report the grade during which they dropped out; for them, the highest grade completed equals the dropout grade minus one. We winsorize the resulting variable such that children can be ahead at most two grades; for example, a 10 year-old child can have completed at most grade six in Kenya and grade five in Tanzania.

The second main outcome variable is the cognitive test score. We construct this score by first standardizing the English, Swahili, and numeracy scores by country, Uwezo survey wave, age, and test booklet to have mean zero and standard deviation one. In a second step, we then average these standardized scores for each student and normalize the resulting composite again to obtain the cognitive test score used in the regressions. In auxiliary regressions, we also use the standardized scores for the individual English, Swahili, and numeracy tests as outcomes.

3. Sample selection

We use data from all available waves of the Uwezo surveys with information on preschool attendance. These are the 2013 and 2014 waves in Kenya and the four waves conducted between 2011 and 2014 in Tanzania. We decided to drop Uganda from the analysis because the only survey with national scope and which collected information on preschool attendance there was fielded in 2013, and information on preschool attendance is missing for 49% of respondents in the corresponding data.

We restrict our attention to children aged 7/8 and above in Kenya/Tanzania because some younger children were still of preschool age at the time of the survey. In order to ensure that we focus on comparable siblings in our within-household analysis, we also drop from the sample any children who report never to have enrolled in preschool or school. Our final sample comprises more than half a million children with information on preschool attendance and at least one of the two main outcomes described above. Note that because a few children are observed with only one of these outcomes, observation numbers in regression tables vary slightly.[8]

References

Funk, C., P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell, and J. Michaelsen. 2015. "The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes." *Scientific Data* 2:150066.

Henderson, J.V., A. Storeygard, and D. Weil. 2012. "Measuring Economic Growth from Outer Space." *American Economic Review* 102:944–1028.

Schady, N., J.R. Behrman, M. Caridad Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, K. Macours, D. Marshall, C. Paxson, and R. Vakis. 2015. "Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries." *Journal of Human Resources* 50:446–463.

[8] All results are robust to focusing on a slightly smaller sample of children observed with both outcomes.

Shah, M., and B.M. Steinberg. 2017. "Drought of Opportunities: Contemporaneous and Long-Term Impacts of Rainfall Shocks on Human Capital." *Journal of Political Economy* 125:527–561.