ONLINE APPENDIX TO:

Preschool Attendance, Schooling, and Cognitive Skills in East Africa

Online Appendix A: Data Appendix

A1. 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

[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 rising difficulty: (1) counting (the number of objects on a show card), (2) recognition of numbers, (3) rank ordering 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.

A2. 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 index of current household wealth, 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

[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.

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 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.

[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/.

Preschool attendance. Recent waves of the Uwezo survey ask respondents whether they ever attended preschool and whether they are currently still enrolled in preschool.[8] 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).

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

The second main outcome variable is the composite 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 score used in the regressions.

In auxiliary regressions, we also use a number of further outcomes. These include an indicator for current enrollment, which takes value 1 if the child reports to be currently enrolled in preschool or school and 0 otherwise. We also construct an indicator for achieving second-grade

[8] The exact questions asked differ slightly across countries and waves. In the 2013 and 2014 waves for Kenya, respondents were asked to indicate whether the child currently attends preschool, with a separate question asking them "How many years of preschool did the child attend?" The 2013 and 2014 waves in Tanzania similarly asked respondents to indicate whether the child currently attends preschool and "If attended, for how many years?" The 2011 and 2012 waves of the survey in Tanzania instead asked "Did you attend preschool (nursery) before joining primary school?"

literacy and numeracy, which takes value 1 if a child achieves the highest competency level in the numeracy test and at least one of the two literacy tests and 0 otherwise.

A3. 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 and above (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.[9]

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

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

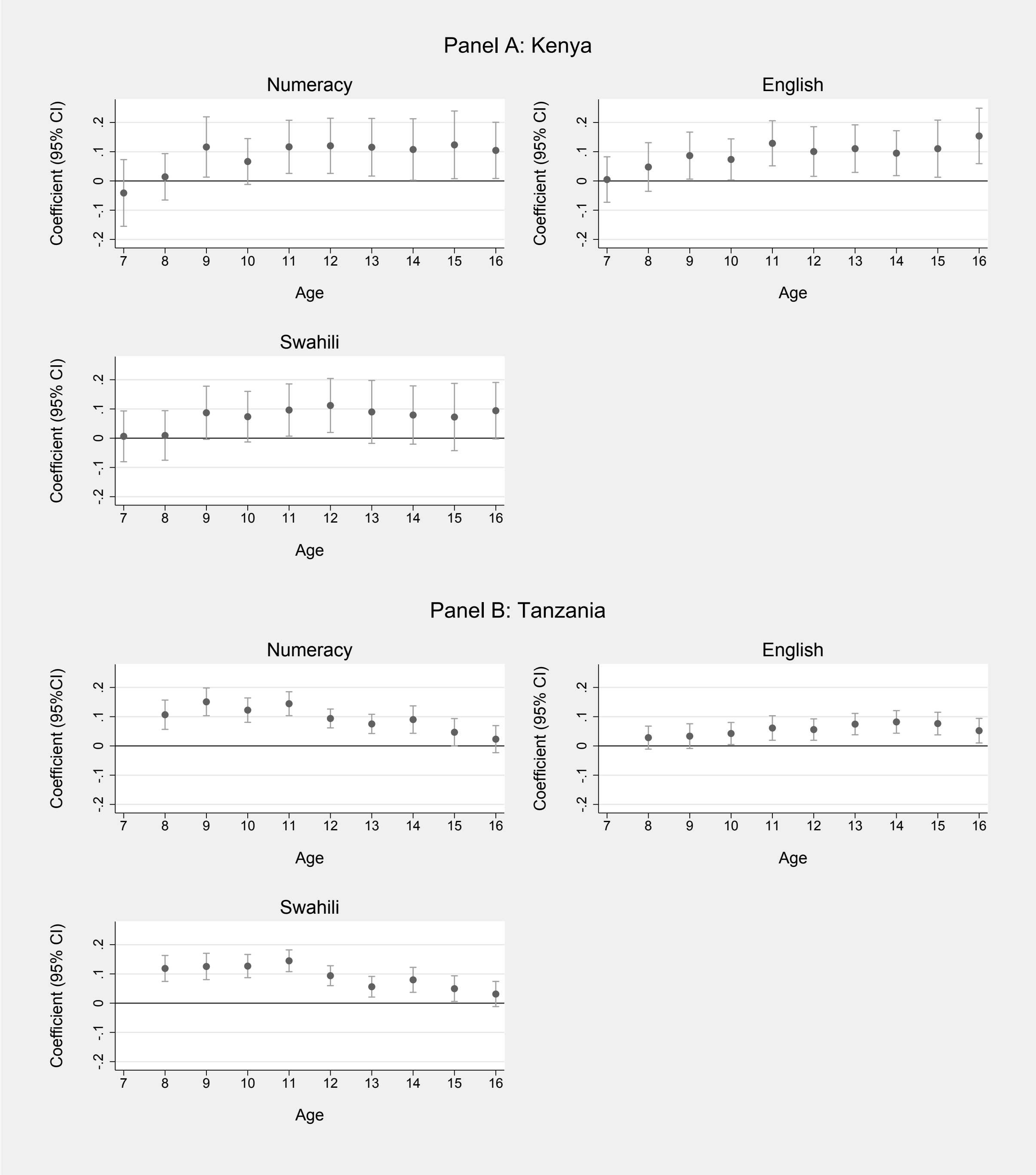
R. Vakis. 2015. “Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries.” *Journal of Human Resources* 50:446–463.

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.

Online Appendix B: Figures and Tables

Online Appendix Figure B.1

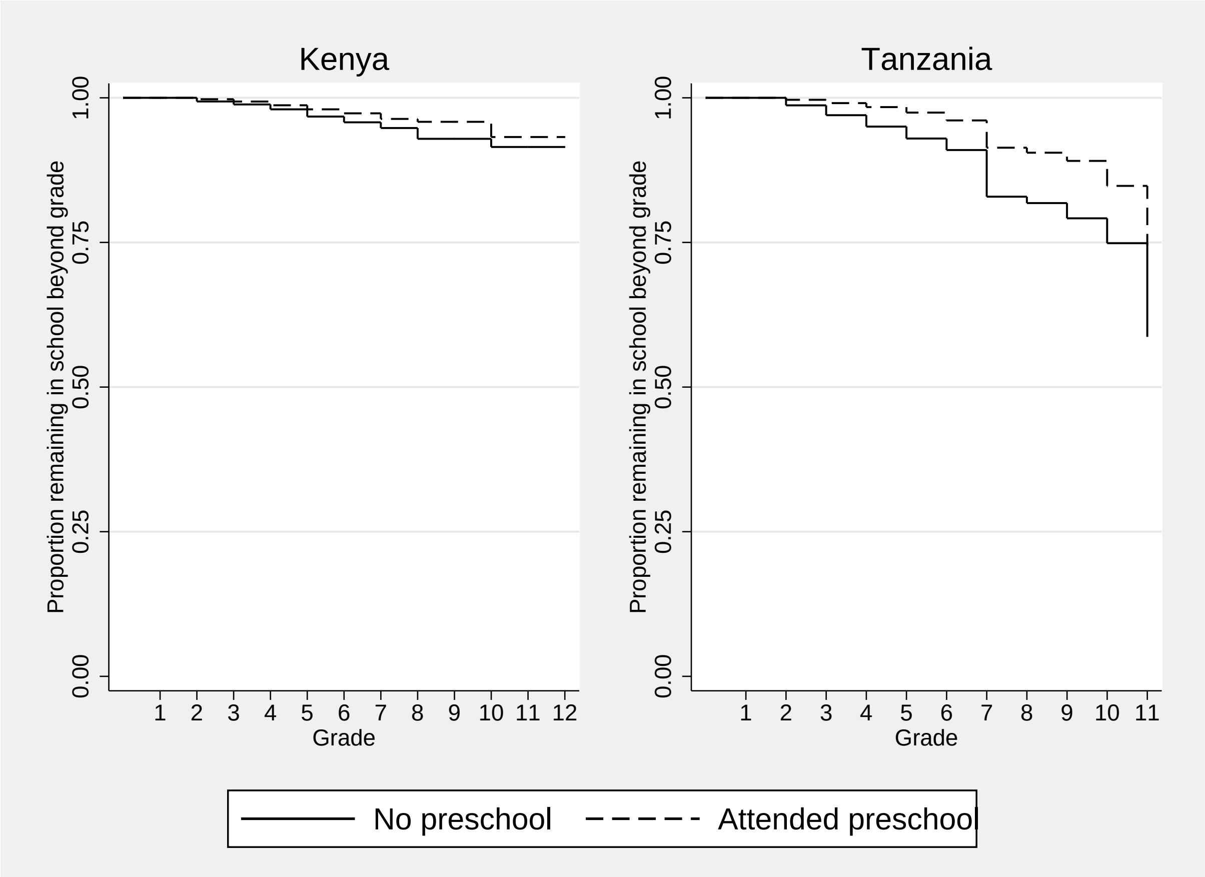
Preschool attendance and numeracy, English, and Swahili skills, by age



Notes: The figure plots coefficient estimates and 95% confidence intervals from regressions of numeracy, English, and Swahili scores on preschool attendance. Scores are standardized by country and age to have mean zero and standard deviation one. The indicator for preschool attendance is interacted with age dummies, and the figure shows the estimated effect of preschool attendance separately for each age. Specifications are otherwise equal to the household fixed effects regressions reported in column 4 of Table 4.

Online Appendix Figure B.2

Kaplan-Meier survival functions for grade progression



Notes: The figure shows Kaplan-Meier survival functions for highest grade attended. The sample is restricted to households with within variation in preschool attendance.

Online Appendix Table B.1

Further robustness checks

	Only siblings born ≤5 years apart		No sampling weights	
	Highest grade attended (1)	Composite test score (2)	Highest grade attended (3)	Composite test score (4)
	Panel A: Kenya			
Attended preschool				
7-9 years old	-0.257***	0.067	-0.377***	0.033
	(0.064)	(0.052)	(0.038)	(0.034)
10-12 years old	-0.083	0.136***	-0.146***	0.124***
	(0.068)	(0.051)	(0.034)	(0.035)
13-16 years old	-0.038	0.123**	-0.026	0.148***
	(0.076)	(0.054)	(0.042)	(0.039)
Observations	158,132	158,408	218,728	218,134
	Panel B: Tanzania			
Attended preschool				
8-9 years old	-0.207***	0.101***	-0.207***	0.103***
	(0.027)	(0.021)	(0.024)	(0.014)
10-12 years old	-0.081**	0.120***	-0.074***	0.112***
	(0.031)	(0.016)	(0.022)	(0.012)
13-16 years old	0.058*	0.077***	0.053**	0.058***
	(0.032)	(0.018)	(0.024)	(0.013)
Observations	235,967	238,852	284,396	288,084

Notes: In columns 1 and 2, the sample is restricted to families with children born at most 5 years apart. Columns 3 and 4 report estimates from regressions that do not use the sampling weights provided with the data. Standard errors in parentheses are clustered at the district level. * p<0.10, ** p<0.05, *** p<0.01.