Program Scale-up and Sustainability

Julie Buhl-Wiggers (Copenhagen Business School)
Jason Kerwin (UMN)
Jeffrey Smith (Wisconsin)
Rebecca Thornton (UIUC)

February 8, 2019

Solving the learning crisis means scaling up interventions

- Primary school enrollment is now very high, but in developing countries children learn very little in school (WDR 2018)
- Huge body of evidence on what works to improve learning (McEwan 2015, Evans & Popova 2016)
- Many roadbloacks to converting evidence into improved education systems:
  - Input quality falls with scale (Allcott 2015, Davis et al. 2017)
  - Implementers vary in quality (Bold et al. 2013, Cameron & Shah 2017)
  - Have to adapt to local conditions (Banerjee et al. 2017)
- Evidence on how best to scale up effective education interventions is limited (but growing)

This Paper

- Data: 5-year panel RCT of a high-impact literacy program in northern Uganda
  - Intervention focuses on mother-tongue-first instruction in grades 1-3
  - Overhauls curriculum, provides detailed teacher guides & lesson plans plus linked textbooks & training
- Experiment embeds a study arm that simulates how programs are often scaled: $\sim 1/3$ the cost, reduces expensive inputs
- Actual scale-up of program occurred in year two of the study
- Follow both students & teachers after intervention to assess how long gains persist
  - Adds to literature on sustained effects of early-childhood interventions (Baird et al. 2015; Gertler et al. 2014; Heckman et al. 2010)

Preview of Results

- Intervention massively improves reading ability: after 3 years, children are 1.35 SDs ahead in local language, 0.73 SDs ahead in English
- High quality and quantity of teacher training and support are crucial for program effects
- Scale-up reduces effectiveness only slightly. Evidence suggests managerial capacity was the issue.
- 50% of student learning gains persist four years after intervention ends
- Treated teachers are still nearly as effective one year later, then impacts drop

The Northern Uganda Literacy Project (NULP)

- Program developed by Mango Tree, a Ugandan education firm
- Two versions: full-cost and reduced-cost
- Full-cost: local language ("Mother Tongue") instruction, detailed lesson plans / scripts, training and monitoring by Mango Tree staff, primers, readers. Runs from Grade 1 to 3.
  - Also provided slates for all students in P1 and clocks in each classroom
- Reduced-cost: Same as full-cost but "cascade" (training-of-trainers) training and monitoring by government staff.
  - Also cut slates and clocks
  - Designed to represent how program could be scaled up

Our data comes from a four-year longitudinal RCT

- RCT was designed to study the impacts of the NULP. Random sample of children tested using EGRA and followed across years.
  - 2013 (38 schools): **Grade 1 (P1)**.
  - 2014 (128 schools): **Grade 1 (P1)**, Grade 2
  - 2015 (128 schools): Grade 1, **Grade 2 (P2)**, Grade 3
  - 2016 (158 schools): Grade 1, Grade 2, **Grade 3 (P3)**, Grade 4

Randomization

- Two waves of schools (2013 and 2014)
  - 2013 schools retained in 2014, program re-started from grade 1
  - Random treatment assignment happened when schools entered study, schools stay in their study arm permanently
- Schools grouped into stratification cells of 3 and randomized by public lottery into one of three arms:
  1. Control group
  2. Reduced-cost NULP
  3. Full-cost NULP
- Two additional features of 2014 randomization:
  1. Cross-randomized provision of slates and clocks to control and reduced-cost schools
  2. One additional school in each stratification cell, excluded from public lottery and testing (pure control)

Four aspects of this study are useful for studying scale-up and sustainability

1. Track one cohort of students that was exposed to treatment only in 2013.
   - Allows us to study fade-out of program effects on students
2. Classrooms & teachers are exposed to treatment when it enters their grade level; we can follow them afterwards
   - Allows us to study fade-out of program effects on *teachers*
3. Reduced-cost treatment designed to simulate how program would be implemented at scale.
4. Actual scale-up of program occurred during experiment, between 2013 and 2014.
   - Program is in P1 in both 2013 and 2014, allowing us to measure effects of scaleup

Our sample includes nearly 31,000 students from 158 schools

	Overall	Control	Full-cost	Reduced-cost	Pure control
Panel A: All students					
# Schools	158	42	42	44	30
# Students	30,966	9,263	9,489	10,168	2,043
# Observations	68,553	21,126	22,232	23,149	2,043
Panel B: Main treated cohort (cohort 2)					
# Schools	158	42	42	44	30
# Students	13,653	3,755	3,838	4,017	2,043
# Observations	35,845	10,814	11,520	11,468	2,043

We observe our main cohort of students every year from 2014-2017.

Student exam score data

- We focus on Early Grade Reading Assessment (EGRA) scores
  - Developed & adapted for local language by RTI
  - Tests various skills needed for reading development, from letter names to word recognition to reading comprehension
  - We use both the English and local language exams

Cohorts and samples of children

- Data for several cohorts of children
  - Cohort 1, treated in 2013 during grade 1 and followed thereafter. In grade 4 during 2016.
  - Cohort 2, treated in 2014-2016 durings grades 1-3. In grade 3 during 2016.
  - Cohorts 3 and 4, not directly treated but in the same schools as treated students. In grades 2 and 1 during 2016.
- Two types of student samples
  1. Initial sample: drawn at beginning of school year, used for balance and to insure against selective attendance/sorting into schools
  2. Top-up sample: selected later during end-of-school exams

Initial sample of students is balanced on observables

	Means			p-value: Identical means across study arms
	Control	Full-cost Program	Reduced-cost Program	
	(1)	(2)	(3)	(4)
Male	0.524	0.514	0.494*	0.167
Age	7.583	7.583	7.555	0.777
Leblango EGRA Reading Index	-0.001	0.011	-0.007	0.734
Letter Name Knowledge (Letters per Minute	1.078	1.241	1.127	0.570
Initial Sound Identification (Sounds Identifie	0.052	0.074	0.061	0.789
Familiar Word Reading (Words per Minute)	0.012	0.021	0.008	0.503
Invented Word Reading (Words per Minute)	0.036	0.013	0.003*	0.242
Oral Reading Fluency (Words per Minute)	0.028	0.051	0.034	0.782
Reading Comp. (Questions Correct)	0.116	0.117	0.112	0.909
Overall				**0.215**

Estimation Strategy

$$Y_{ist} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{\gamma_s}' + u_{ist}$$

Y_{ist}: test scores for student i in school s at the end of year t

- Use PCA indices across scores to avoid multiple comparisons
- Typically present results in SDs of control-group distribution

$\boldsymbol{\gamma_s}$: vector of stratification cell indicators

$FullCost_s$ and $ReducedCost_s$ are treatment indicators for school s

Estimation Strategy

$$Y_{ist} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{\gamma_s}' + u_{ist}$$

Y_{ist}: test scores for student i in school s at the end of year t

- Use PCA indices across scores to avoid multiple comparisons
- Typically present results in SDs of control-group distribution

γ_s: vector of stratification cell indicators

$FullCost_s$ and $ReducedCost_s$ are treatment indicators for school s

Main specification was laid out in pre-registered analysis plan.

Estimation Strategy

$$Y_{ist} = \beta_0 + \beta_1 FullCost_s + \beta_2 ReducedCost_s + \boldsymbol{\gamma_s}' + u_{ist}$$

Y_{ist}: test scores for student i in school s at the end of year t

- Use PCA indices across scores to avoid multiple comparisons
- Typically present results in SDs of control-group distribution

$\boldsymbol{\gamma_s}$: vector of stratification cell indicators
$FullCost_s$ and $ReducedCost_s$ are treatment indicators for school s

Main specification was laid out in pre-registered analysis plan.

Cluster SEs by school (level of treatment). When number of schools is small, check robustness to randomization inference.

Full-cost NULP sharply improves mother-tongue reading by end of Grade 3

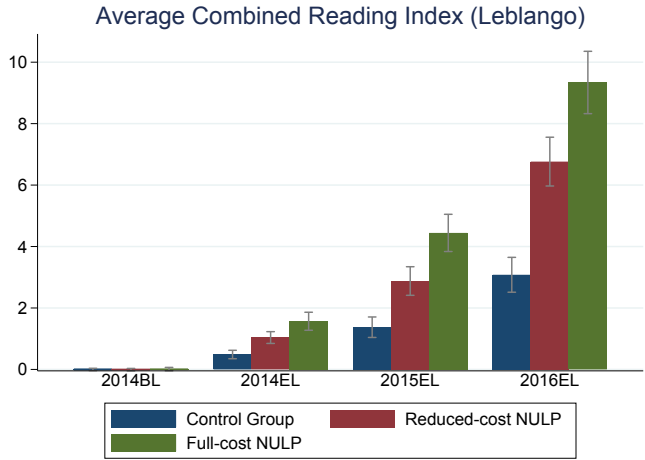
	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Name Recognition (letters/minute)		Oral Reading Fluency (words/minute)		Combined Reading Index (grade level equivalents)	
	Score	SDs	Score	SDs	Score	SDs
Full-cost Program	22.164***	1.431***	12.563***	1.180***	6.242***	1.348***
	(1.552)	(0.100)	(1.044)	(0.098)	(0.495)	(0.107)
Reduced-cost Program	13.238***	0.855***	7.140***	0.671***	3.627***	0.784***
	(1.392)	(0.090)	(0.999)	(0.094)	(0.453)	(0.098)
Difference between full-cost and reduced-cost treatment	8.926***	0.576***	5.423***	0.510***	2.614***	0.565***
	(1.619)	(0.104)	(1.175)	(0.110)	(0.526)	(0.114)
Control Group Mean	17.922	0.000	5.327	0.000	3.081	0.000
Control Group SD	15.492	1.000	10.643	1.000	4.629	1.000

Large impacts on English reading ability as well

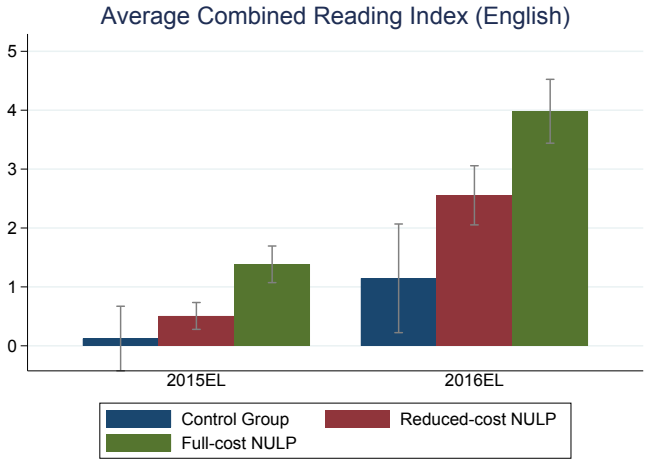
	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Name Recognition (letters/minute)		Oral Reading Fluency (words/minute)		Combined Reading Index (grade level equivalents)	
	Score	SDs	Score	SDs	Score	SDs
Full-cost Program	1.514	0.083	5.127***	0.280***	2.806***	0.729***
	(1.231)	(0.067)	(1.615)	(0.088)	(0.380)	(0.099)
Reduced-cost Program	1.126	0.061	2.226	0.121	1.551***	0.403***
	(1.207)	(0.066)	(1.401)	(0.076)	(0.331)	(0.086)
Difference between full-cost and reduced-cost treatment	0.388	0.021	2.900**	0.158**	1.255***	0.326***
	(1.162)	(0.063)	(1.206)	(0.066)	(0.315)	(0.082)
Control Group Mean	13.263	0.000	8.371	0.000	1.145	0.000
Control Group SD	18.347	1.000	18.342	1.000	3.851	1.000

Among the largest-ever gains ever for a primary-school intervention (McEwan 2015)

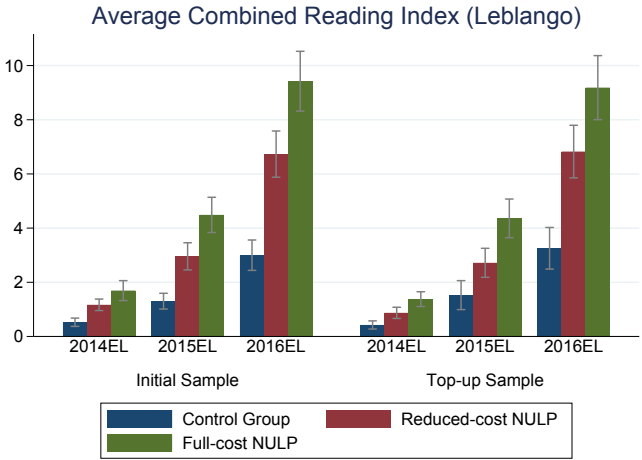
Learning gains build over grades 1-3



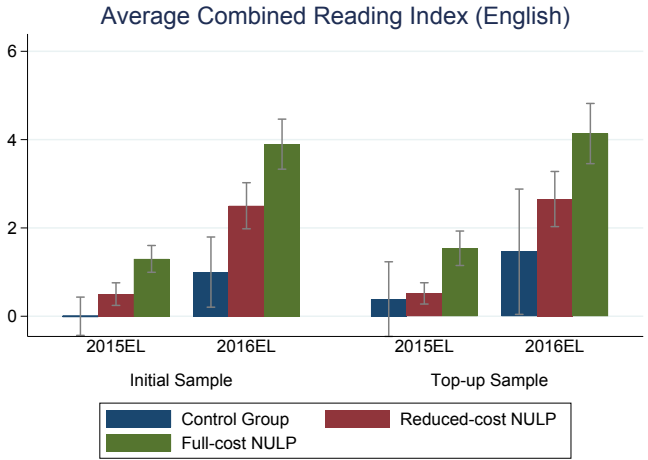
English scores are measured in grades 2 and 3



Initial vs. top-up sample does not matter for results



No evidence that students select into treatment schools



Hawthorne effects?

- Potential concern: just interacting with these schools might change outcomes
- Impacts could be overstated:
  - Repeated testing of control schools could induce fatigue & low effort
  - Interactions with implementer could also increase effort *per se*
- Or they could be understated:
  - Control group received small gifts from implementers (chalk, wall charts) to encourage participation
- We held out one school per stratification cell in 2014 to test for these issues
  - These 30 "pure control" schools were only tested in 2016

Nearly-identical outcomes in pure control & control schools

	(1)	(2)	(3)	(4)
	Mother-Tongue Reading Index (grade level equivalents)		English Reading Index (grade level equivalents)	
	Raw Score	SDs	Raw Score	SDs
Full-cost Program	6.573***	1.512***	3.184***	1.039***
	(0.507)	(0.117)	(0.305)	(0.099)
Reduced-cost Program	3.967***	0.913***	1.871***	0.610***
	(0.504)	(0.116)	(0.349)	(0.114)
Pure Control	0.020	0.005	-0.383	-0.125
	(0.305)	(0.070)	(0.283)	(0.092)
Control Group Mean	2.852	0.000	0.630	0.000
Control Group SD	4.346	1.000	3.064	1.000

How do we get these learning gains to as many students as possible?

Given these major improvements in learning, the next question is how we can expand the program and sustain its impacts.

Examine this question two different ways:

1. Estimate effect of reduced-cost version of program that simulates how program might be scaled up
2. Study actual scale-up of program between 2013 and 2014

Reduced-cost program has sharply lower impacts

	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Name Recognition (letters/minute)		Oral Reading Fluency (words/minute)		Combined Reading Index (grade level equivalents)	
	Score	SDs	Score	SDs	Score	SDs
Full-cost Program	22.164***	1.431***	12.563***	1.180***	6.242***	1.348***
	(1.552)	(0.100)	(1.044)	(0.098)	(0.495)	(0.107)
Reduced-cost Program	13.238***	0.855***	7.140***	0.671***	3.627***	0.784***
	(1.392)	(0.090)	(0.999)	(0.094)	(0.453)	(0.098)
Difference between full-cost and reduced-cost treatment	8.926***	0.576***	5.423***	0.510***	2.614***	0.565***
	(1.619)	(0.104)	(1.175)	(0.110)	(0.526)	(0.114)
Control Group Mean	17.922	0.000	5.327	0.000	3.081	0.000
Control Group SD	15.492	1.000	10.643	1.000	4.629	1.000

Less effective at raising English scores as well

	(1)	(2)	(3)	(4)	(5)	(6)
	Letter Name Recognition (letters/minute)		Oral Reading Fluency (words/minute)		Combined Reading Index (grade level equivalents)	
	Score	SDs	Score	SDs	Score	SDs
Full-cost Program	1.514	0.083	5.127***	0.280***	2.806***	0.729***
	(1.231)	(0.067)	(1.615)	(0.088)	(0.380)	(0.099)
Reduced-cost Program	1.126	0.061	2.226	0.121	1.551***	0.403***
	(1.207)	(0.066)	(1.401)	(0.076)	(0.331)	(0.086)
Difference between full-cost and reduced-cost treatment	0.388	0.021	2.900**	0.158**	1.255***	0.326***
	(1.162)	(0.063)	(1.206)	(0.066)	(0.315)	(0.082)
Control Group Mean	13.263	0.000	8.371	0.000	1.145	0.000
Control Group SD	18.347	1.000	18.342	1.000	3.851	1.000

Is the reduced-cost version more cost-effective?

Tentative results, using costs from 2013:

- MC/student is $15.39/year for full-cost program, $6.05/year for reduced-cost
- Both variants raise scores by about 0.02 SD/dollar in English
- For mother tongue, reduced-cost program returns 0.04 SD/$, full-cost returns 0.03 SD/$

However: reduced-cost version hurts writing scores in P1 (Kerwin and Thornton 2018)

- And cost-effectiveness is highly sensitive to which outcome measure we pick

Estimated cost difference is an upper bound

- Full-cost program costs most in P1 — no slates in P2 & P3

Differences in materials don't explain the gap in outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Mother Tongue			English		
	Oral Reading Fluency	Reading Comp.	Combined Reading Index	Oral Reading Fluency	Reading Comp.	Combined Reading Index
Full-cost Program	1.220***	1.018***	1.478***	0.421***	0.340***	0.854***
	(0.152)	(0.124)	(0.165)	(0.0797)	(0.0689)	(0.109)
Reduced-cost Program						
With both slates and clock	0.426*	0.468***	0.572***	0.122	0.0693	0.259*
	(0.217)	(0.157)	(0.218)	(0.128)	(0.132)	(0.156)
With slates only	0.682***	0.608***	0.897***	0.148	0.180	0.487***
	(0.226)	(0.179)	(0.237)	(0.129)	(0.115)	(0.174)
With clocks only	0.903***	0.833***	1.136***	0.312***	0.186**	0.600***
	(0.155)	(0.132)	(0.171)	(0.0905)	(0.0813)	(0.116)
Neither slates nor clocks	0.771***	0.733***	0.981***	0.415***	0.356***	0.688***
	(0.231)	(0.186)	(0.239)	(0.127)	(0.104)	(0.157)

Differences in outcomes driven by quantity & quality of training & support

- Both treatment groups identical on
  - Instructional philosophy
  - Emphasis on mother-tongue instruction (Kerwin & Thornton 2018)
  - Teacher guides & lesson plans
  - Textbooks
  - Training content
- Reduced-cost program differs in two ways
  - Some schools didn't have certain materials (doesn't matter)
  - Delivery of training & support

Cascade training models and cost-cutting

- NULP training is expensive
  - Offsite training w/teaching experts 4X/year + intensive support
  - At least 50% of the gap in costs between full- and reduced-cost is due to training
- Reduced-cost model used "cascade"/"train-the-trainers" strategy to cut costs:
  - In particular, utilizing existing education department staff
  - Common approach — e.g. the School Health and Reading Program (RTI 2016)
- Also scaled back check-up visits to support teachers & give feedback
  - From 15/year to 6/year
- These cost-cutting measures significantly reduce impacts

What happens when the program actually scales up?

- After initial year of the study, we secured funding to expand sample of schools
  - From 38 schools (26 treated) to 128 schools (86 treated)
  - Had to relax school eligibility criteria to achieve this
- In both years, schools had to:
  - Have desks and blackboards in P1 classrooms
  - Be accessible by road year-round
  - Not have previously received Mango Tree support

Program expansion led to lower school eligibility criteria

- In 2013, imposed the following additional restrictions:
  1. Two P1 classrooms & teachers
  2. Lockable cabinets
  3. head teacher regarded as "engaged" by CCT
  4. $\leq$ 135 students/teacher
  5. School must be $\leq$ 20km from CC
- For the additional schools in 2014:
  - Restrictions 1-3 were dropped
  - Restriction 4 was relaxed to a cutoff of 150 students/teacher
  - Restriction 5 was relaxed to a maximum distance of 22km

Scale-up slightly reduced the gains in original schools

	(1)	(2)	(3)	(4)	(5)	(6)
	Mother Tongue Letter Name Recognition			Mother Tongue Combined Reading Index		
	2013	2014 (86 Treated Schools)		2013	2014 (86 Treated Schools)	
	(26 Treated Schools)	Original Schools	New Schools	(26 Treated Schools)	Original Schools	New Schools
Full-cost Program	1.043***	1.046***	1.112***	0.824***	0.610***	0.828***
	(0.163)	(0.244)	(0.132)	(0.147)	(0.193)	(0.115)
Reduced-cost Program	0.418**	0.674***	0.713***	0.156	0.233	0.467***
	(0.181)	(0.219)	(0.115)	(0.122)	(0.165)	(0.101)
Observations	1,476	1,081	4,527	1,460	1,070	4,490
Number of Schools	38	38	90	38	38	90

Managerial capacity and input quality

- Expansion of program appears to have slightly strained managerial capacity
  - Somewhat lower gains in original schools
  - NGO had to hire more implementing staff & managers
  - Potentially selecting from a less-experienced group (Davis et al. 2017)
  - Alternatively: could be original P1 teachers losing some enthusiasm
- If anything, quality of other inputs went *up*
  - Gains in new schools are higher than those for original schools
  - Arguably a lower bound on input quality — management capacity was strained
  - This is the opposite of the pattern documented in Allcott (2015)

Sustainability and program scale-up

Two major concerns with scaling this program up:

1. Common cost-cutting techniques reduce the effectiveness of the program
2. Scaling up program as-is can strain managerial capacity/hit labor constraints

If gains are sustained, maybe we can work around these problems:

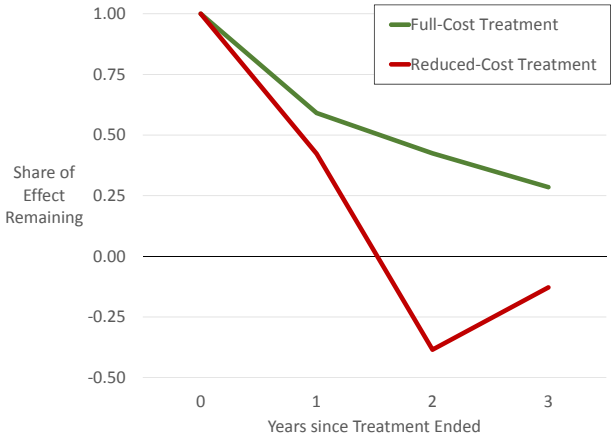
- Imagine an intervention that permanently improves a teacher's quality
- Suppose you only have the capacity to intervene in $\sim 10\%$ of schools at a time
- Over 10 years, can scale up to all schools without running into usual constraints

To that end, we also examine how long the NULP's impacts persist.

How long do learning gains persist?

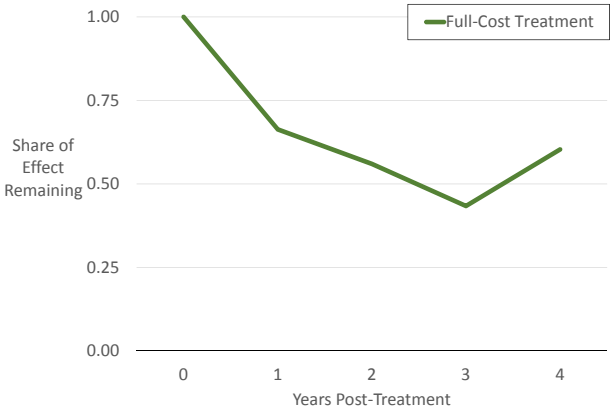
- Follow cohort of students who were treated as first-graders for the next four years
  - Test changed in 2017, dropping some subtests; can do combined scores only until P4
- Compute treatment effects for each year in *contemporaneous* control-group SDs
  - E.g. in P2, treatment effects in SDs of control-group P2 outcomes
- Divide each year's treatment effect by effect for P1
- Similar process for treated *classrooms*
  - Grade levels in a school that got treatment in a previous year
  - For treated *teachers*, track whether teacher that received training is still around

Overall student gains decay by 20% per year



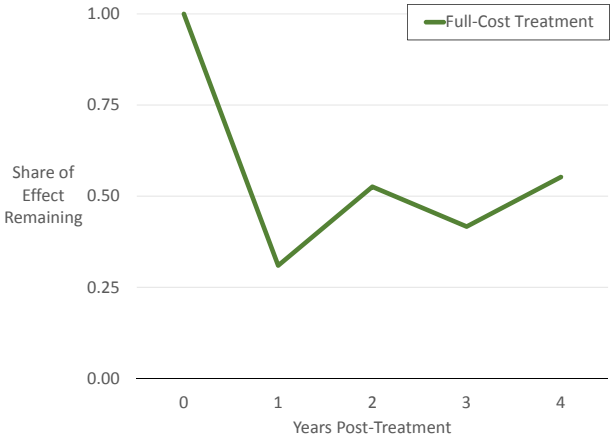
Substantially faster drop & smaller initial gains for reduced-cost $\Longrightarrow$ focus on full-cost

Oral reading fluency gains persist for longer

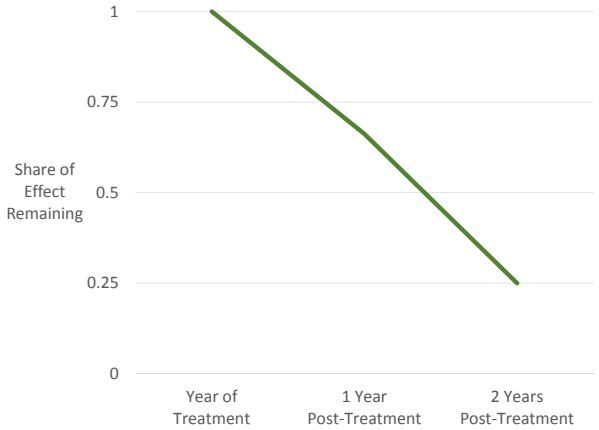


Rate of decline is about 10% per year

Reading comprehension gains are still 0.25 SDs, four years post-treatment



How long do effects on treated P1 *classrooms* last?



Most classroom gains fade out within two years.

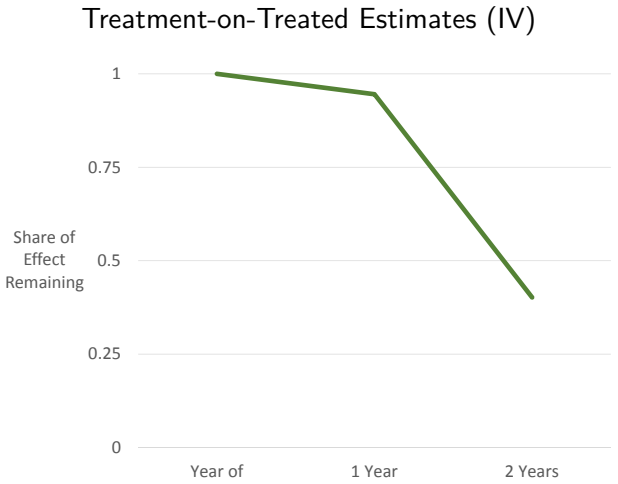
Many teachers leave classrooms within a few years of treatment ending

Share of Treated Teachers Still in Same School & Grade

	(1) Year of Treatment	(2) 1 Year Post-Treatment	(3) 2 Years Post-Treatment
P1	2014	2015	2016
Full-cost Program	1.00	0.94	0.84
Reduced-cost Program	1.00	0.87	0.84
P2	2015	2016	
Full-cost Program	1.00	0.68	
Reduced-cost Program	1.00	0.48	

Fadeout possibly due to teacher attrition, but also forgetting, loss of motivation, etc.

Gains persist longer if we focus on treated P1 teachers



Which inputs prevent scaleup from succeeding?

- Quality & quantity of training is key bottleneck to successful program scale-up
  - Even at small scale, a cascade training model was much less effective
- Supply of managerial capacity is fairly elastic in our context
  - Quadrupling number of treated schools led to at most modest declines in impacts
- Implementers better at in selecting own staff than other inputs (e.g. schools)?
  - Original schools selected for ease of implementation
  - But new schools, w/**worse** physical inputs & lower staff numbers, had **bigger** gains
  - Marginal product is increasing rather than decreasing

Achieving cost-effective scale-up

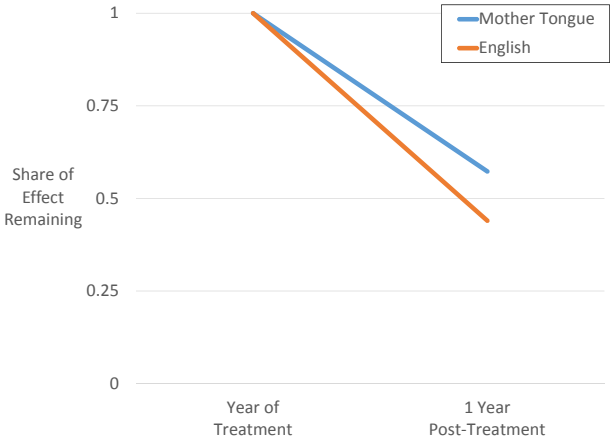
High-impact education interventions can have long-lasting benefits

- Teachers retain over 90% of gains one year post-intervention
  - Instead of cutting costs by lowering training quality, alternate years of training?
  - Or instead of repeating training, some other support to help sustain gains?
- Student learning gains persist in the long term
  - But only if the intervention is strong enough — not if it is watered down
  - Costlier program looks more cost-effective for scaling up at longer time scales

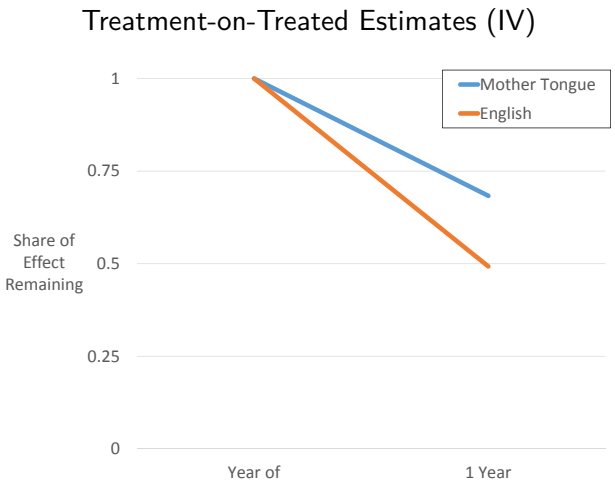
- Thank you!
- Please contact me if you have any other questions or comments:
  jkerwin@umn.edu
  www.jasonkerwin.com

Bonus Slides

Classroom-level treatment effect persistence for P2



Teacher-level treatment effect persistence for P2



Grade 4: Partial Project Phase-Out

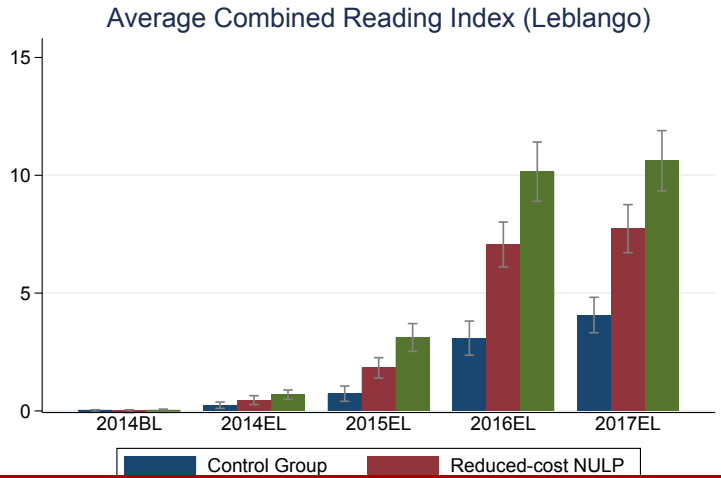
- Original plans called for program implementation in grades 1-3
- Main treated cohort of students entered grade 4 in 2017
- During 2017: NGO split off of Mango Tree parent company, management changed
- Some materials development (textbooks/teacher guides) for grade 4, treated schools received some intervention but not much

Implementation was weak in 2017

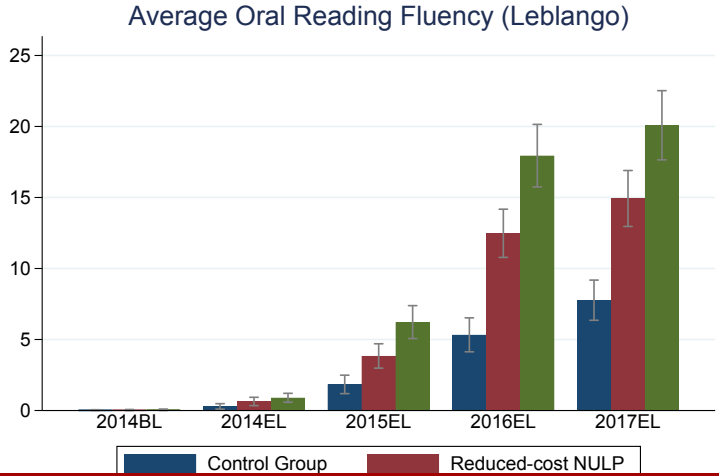
Classroom Support Supervision Visits in 2017

	(1) Mango Tree Staff Visits	(2) CCT Visits	(3) Total Visits
Full-cost Program			
Total Scheduled	9	6	15
Share Completed	0.06	0.15	0.10
Reduced-cost Program			
Total Scheduled	0	6	6
Share Completed	-	0.58	0.58

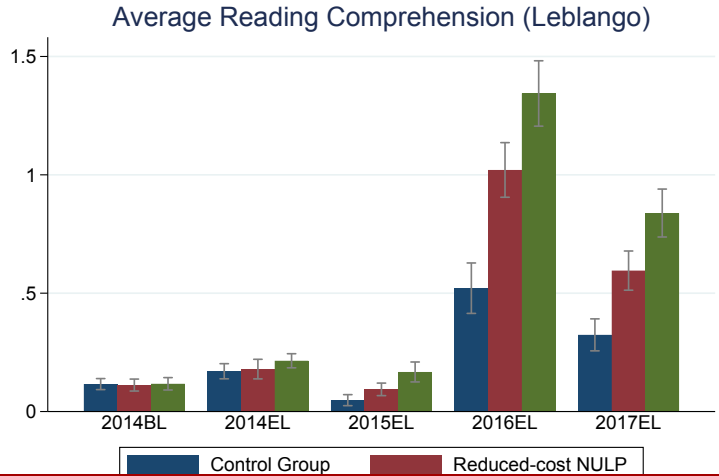
2014-2017 Results — Mother-Tongue Overall Reading



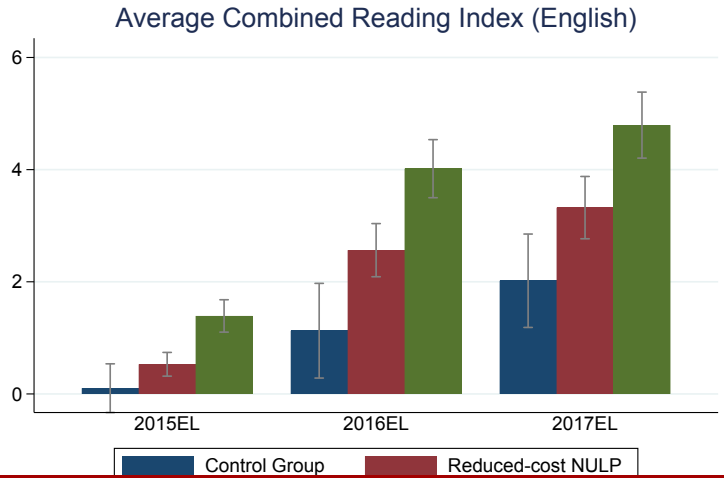
2014-2017 Results — Mother-Tongue Reading Fluency



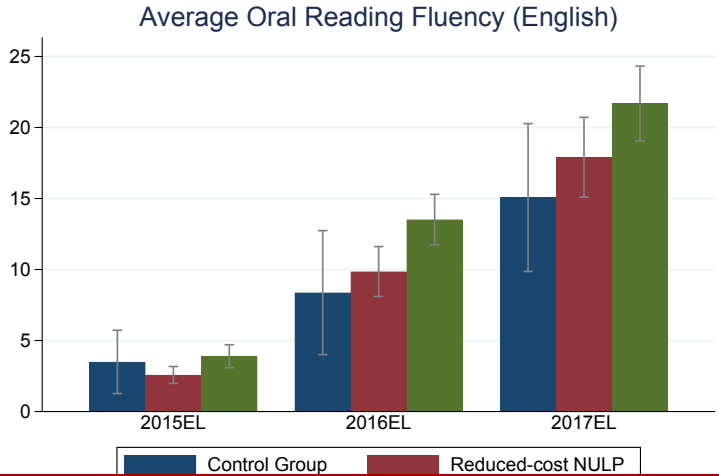
2014-2017 Results — Mother-Tongue Reading Comp.



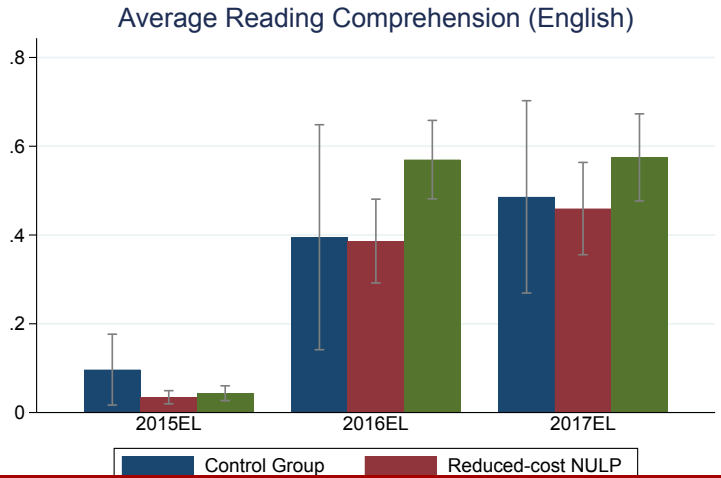
2014-2017 Results — English Overall Reading



2014-2017 Results — English Reading Fluency



2014-2017 Results — English Reading Comp.



2017 Results: Small Treatment Effects or Strong Persistence?

- If we consider 2017 as an untreated year, it is the first period we can observe students who have been through the full program (P1-P3)
  - Effects are strongly persistent - treatment-control gaps remain on all major outcomes
- If instead 2017 was a treated year, the treatment was very weak
  - Virtually no increase in treatment-control score gap
- Reality is probably between the two extremes: students got a weak treatment but most of the score gap is just persistence
  - Future work: process & digitize documentation about what was done in each school in 2017