BREAD-IGC Virtual Development Course

Education: Technology in Education

Karthik Muralidharan

UC San Diego, NBER, BREAD, and J-PAL

11 February 2022

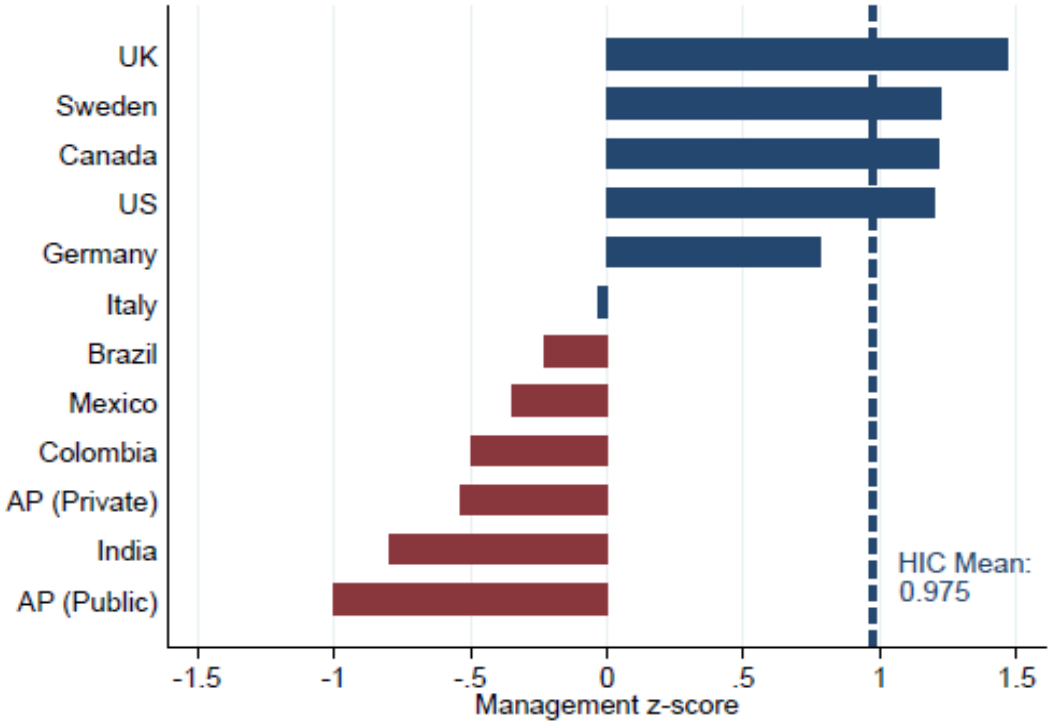
School Governance Wrap-up

- Teacher absence
- Monitoring: top-down & bottom-up
- Teacher pay and teacher incentives
- Contract teachers
- **School management**

Lemos, Muralidharan, and Scur(2021)

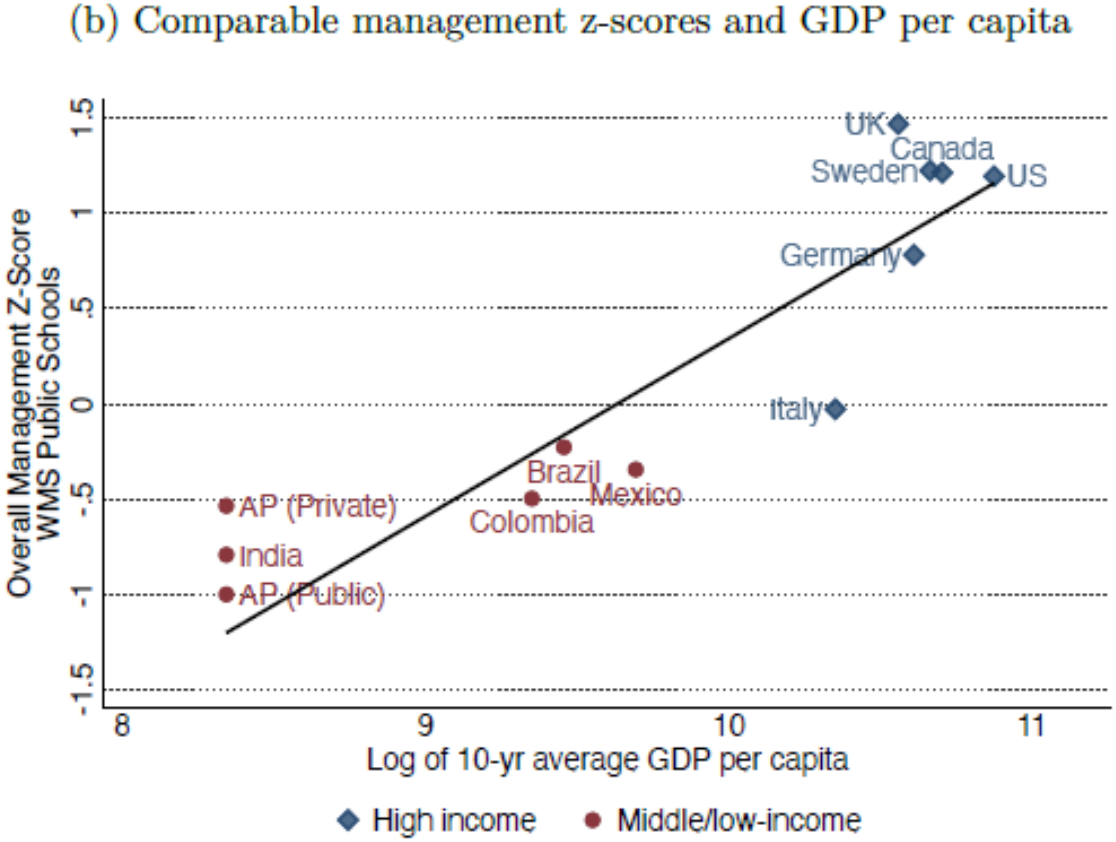
Figure 1: Global benchmarks

(a) Rank of comparable management z-scores



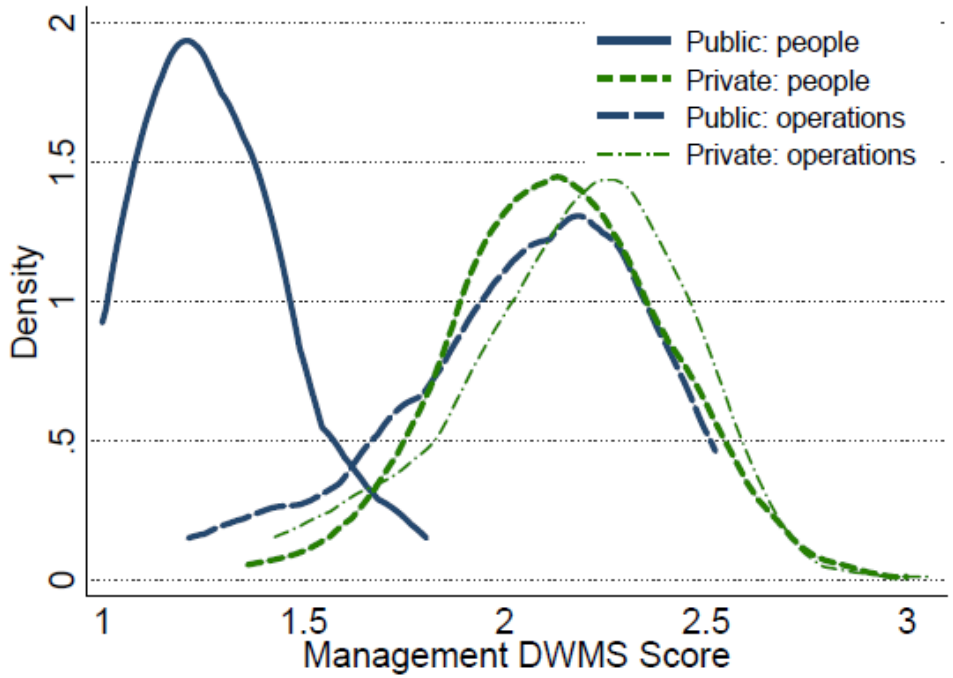
- Develop and use the D-WMS to be comparable to the WMS but much more granular to allow us to better capture finer variation in the left tail
- School management scores are systematically lower in low-income countries

Lemos, Muralidharan, and Scur(2021)



- Not an outlier after adjusting for income – but suggests that variation in TFP in education systems may also partly explain income-outcome gradient
- Thus, improving quality of school management may be an important component of improving developing country education systems

Lemos, Muralidharan, and Scur(2021)



- Public schools do especially poorly on personnel management
- Consistent with differences in absence, active teaching, etc. observed in the data
- Also consistent with our finding a strong positive correlation between teacher value-added and pay in private schools, and no such correlation in public schools
- Management scores are also correlated with teaching activity & value added

Can we improve public school management at scale?

- LMIC Governments increasingly understand the importance of school management and are keen to improve it
  - Many such reforms (160 examples in 84 countries in WB database)
  - But very little evidence on impact

- Muralidharan & Singh (2021): Large-scale RCT of a school management intervention in the Indian state of MP

- The program (MPSQA) was modelled after several global 'best practices' in school management (and developed with several international expert inputs)
  - Comprehensive school quality assessments
  - Customized school improvement plans
  - Regular (intended) follow up by cluster-resource coordinators
  - However, no incentives (explicit or implicit)

- Randomized and evaluated at scale with 1,774 treatment schools

Results

Four main results:

1. **Assessments were completed and of high quality**
   - ~93% of elementary schools assigned to treatment group were covered by the assessments and had SIPs prepared
   - Contain meaningful variation across schools
   - Most schools rated inadequate (collusion unlikely)
2. **No improvement in support or oversight in treated schools**
   - No change in frequency of visits or the content of inspections
   - School Management Committees also did not play a more active role in treatment schools
3. **No change in teacher effort and classroom processes**
   - Teacher absence was high (33%) and unchanged
   - No impact on instructional time, use of textbooks and workbooks, or the likelihood of checking student homework books
   - Student absence rates were also high (47%) and unaffected
4. **Also, no change in test scores**
   - 18 months after the assessments
   - Across administrative tests or our own

Evaluating a further scale-up

- The government had already planned for the expansion of the program to the next phase of ∼25,000 schools in late 2016
- Additional impetus from a similar national program
  - External assessments were replaced by school self-assessments
  - The plans were made much more detailed
  - No change in incentives
  - Already scaled up to >600k schools, targets 1.6 million schools
- We evaluate this scale up using a matched-pair design (satisfying parallel trends in previous years) in 10 districts
- We again find no evidence of improved learning outcomes 1.5 years after the program was rolled out

Understanding program implementation and failure

- We collected extensive qualitative data from teachers, head-teachers and education officials
- We document that, for these officials, the program was reduced to an exercise in administrative compliance
  - Both teachers and supervisors perceived the program primarily as a data collection and paperwork filling effort
  - Paperwork was submitted on time.
  - Program delivery effectively ceased after filing SIPs
  - De facto, the reform was very far from the reflective exercise in self-evaluation and improvement envisaged in the program design.
- These features relate much more broadly to bureaucratic incentives in the public sector that mainly reward the *appearance of activity* rather than improving outcomes (which are not measured)

Contributions 1/2

Improving management quality in developing countries

- Improving management quality in developing countries
  - Management quality is correlated with productivity in both the private and public sectors (Bloom and Van Reenen 2007, Rasul and Rogger 2018, Rasul et al 2018)
  - Providing management consulting inputs leads to persistent improvements in productivity in private firms in India and Mexico (Bloom et al 2013, Bruhn et al 2018)
- We provide experimental evidence on the impact of an attempt to improve management quality in the **public sector**
  - No impact on either processes or outcomes
  - Consistent with other RCT evidence that inputs that are effective in private schools may have no impact in public schools (in LMICs)
- Conjecture that important reason is the lack of incentives
  - Growing evidence of complementarities across inputs/knowledge and incentives in education (Mbiti et al. 2019, MS 2011), health (Das et al. 2016), and even private sector (Atkin et al 2017)
  - Contrast with results from US, UK where school ratings did have an impact when combined with threat of sanctions for low-performing schools and principals (Figlio and Loeb 2011; Rockoff and Turner 2010; Hussain 2015)

Contributions 2/2

Organizational Economics and Public Sector Reform

- Shows the difficulties of "change management" in large organizations
  - Large and active literature, with several theories and case studies, but very little well-identified evidence (By, 2005)
  - Importance of well-identified null results (Abadie 2020)
  - Null results may also reflect intensity of the intervention
- Illustrate the nature of bureaucratic incentives (and divergence between actual and perceived 'success')
  - The program was deemed a success by administrative metrics
  - We show that the program's success was only on paper
- Striking example of institutional isomorphism (DiMaggio and Powell 1983, Pritchett 2013)
  - Such isomorphic mimicry may help to explain both the initial enthusiasm for the program and the subsequent scale up
- Highlights the importance of independent evaluations of development projects and programs

Bureaucratic incentives under weak state capacity

What stands out here are higher-level officials in the administrative hierarchy making decisions about programs and targets that bear little relevance to realities on the ground; also present, in turn, are ***subordinates faithfully executing programs on paper but caring little for how well they are implemented.*** *Targets are indeed met, but the ultimate goals of the programs go unfulfilled.*

Gupta (2012)
Red Tape: Bureaucracy, Poverty and Structural Violence

Why do governments continue scaling these up?

Institutional isomorphism

One compelling explanation is provided by the theory of "institutional isomorphism"

> *"Organizations tend to model themselves after similar organisations in their field that they perceive to be more legitimate or successful," (p. 152);*
>
> *"these institutional isomorphic processes can be expected to proceed [even] in the absence of evidence that they increase organisational efficiency," (p.153);*
>
> *and that such mimicry has "a ritual aspect; [organizations] adopt these "innovations" to enhance their legitimacy, to demonstrate they are at least trying to improve." (p. 151)*

DiMaggio and Powell (1983)

Conclusions – What should we take away from all of this?

- Returns to improving management quality may be especially high in public sector
  - But it may also be *much* more difficult given nature of bureaucratic incentives
- We cannot identify what factors *would* lead to success
- But 3 factors seem quite important (from other studies of interventions that have been found to work)
  - Better incentives for improving effort or outcomes
  - Better visibility on outcomes at the beneficiary level
  - Additional staffing
- Governments and donors are constantly designing and deploying programs to improve service delivery and development outcomes
  - Programs are often judged on design quality ("best practices") and number of people reached
  - On these measures, the program was a resounding success!
  - Highlights importance of independent evaluations of impact

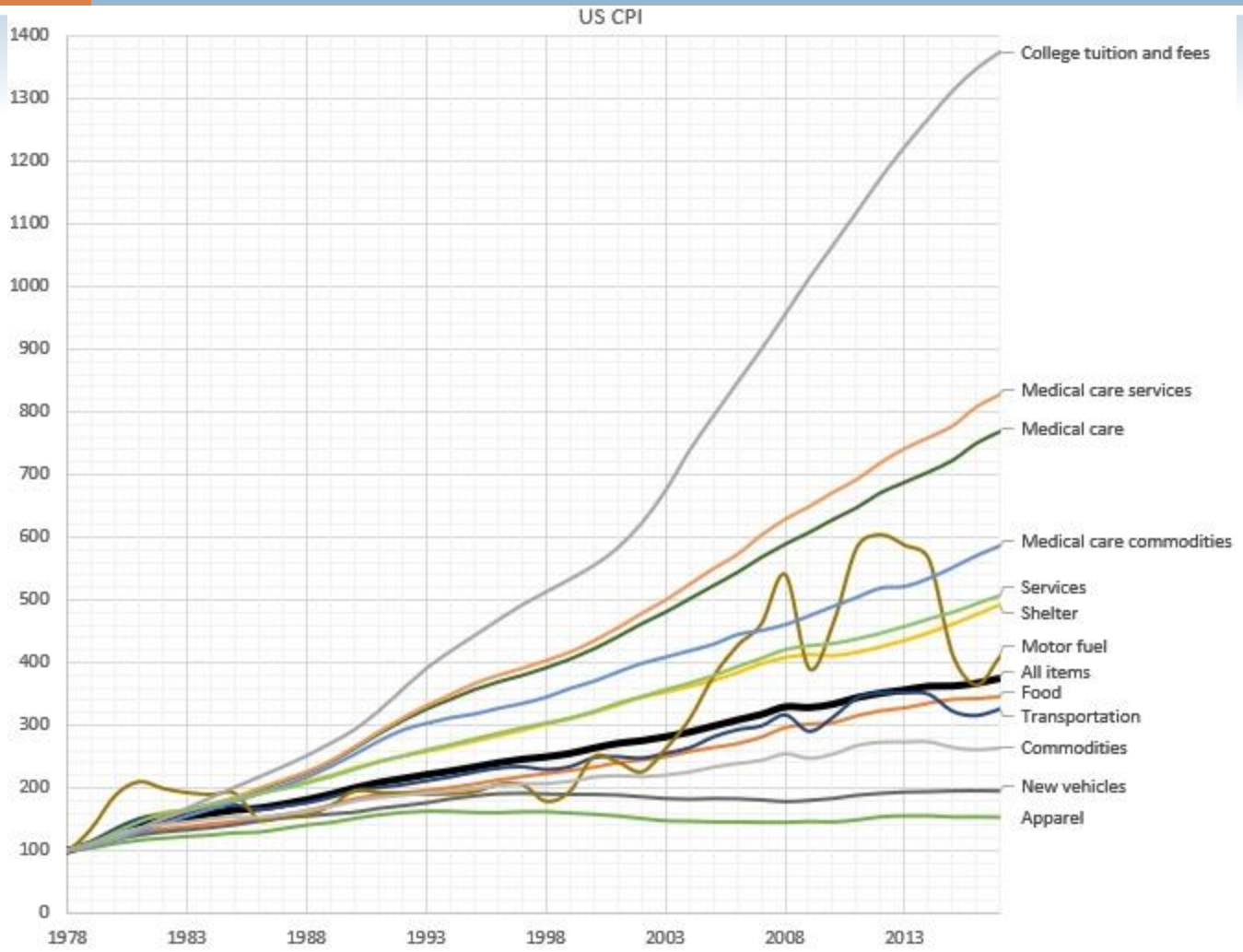
Technology in Education

- Productivity gains in education production have been limited compared to rest of the economy (around the world)
  - Seen in higher cost inflation relative to CPI
  - Case of Baumol's disease – education & health (education may be worse)
  - Technology of instruction has basically been unchanged (chalk & talk)
  - Hence, lots of excitement about technology in education/instruction

- Mechanisms of potential impact include
  - Cost-effective access to high-quality instruction
  - Leapfrog constraints in teacher knowledge
  - Supplemental instruction, practice, reinforcement at home
  - Customizing learning paths for students (and inducing greater engagement)
  - Shortening feedback look
  - Better engagement with parents

- Technology can also be used to improve governance include
  - Monitoring teacher attendance; improving outcome measurement

Baumol's Disease Illustrated



However evidence on CAL has been mixed & not lived up to hype

- Banerjee et al (2007)
  - Large positive effects (0.47SD) of a CAL program run by Pratham
  - However, found to be 5-7 times less cost effective than using balsakhis (similar demographic as the contract teachers)

- Cristia et al (2017) & Beuermann et al (2015)
  - RCT of the impact of school and home use of the XO Laptop (OLPC)
  - Positive impacts on knowledge of computer use
  - But no impact on measures of cognitive development

- Malamud & Pop-Eleches (2011)
  - RD-based study of providing home computers to middle-schoolers in Romania
  - Improved computer schools
  - But *hurt* grades on core school subjects (Math, English, Romanian)
  - Most students report playing computer games on a daily basis (!); suggestive evidence of reduced time reading and doing homework

- Overall, the evidence points to: “mixed evidence with a pattern of null effects” (Bulman and Fairlie 2016 review article)

The *Mindspark* program

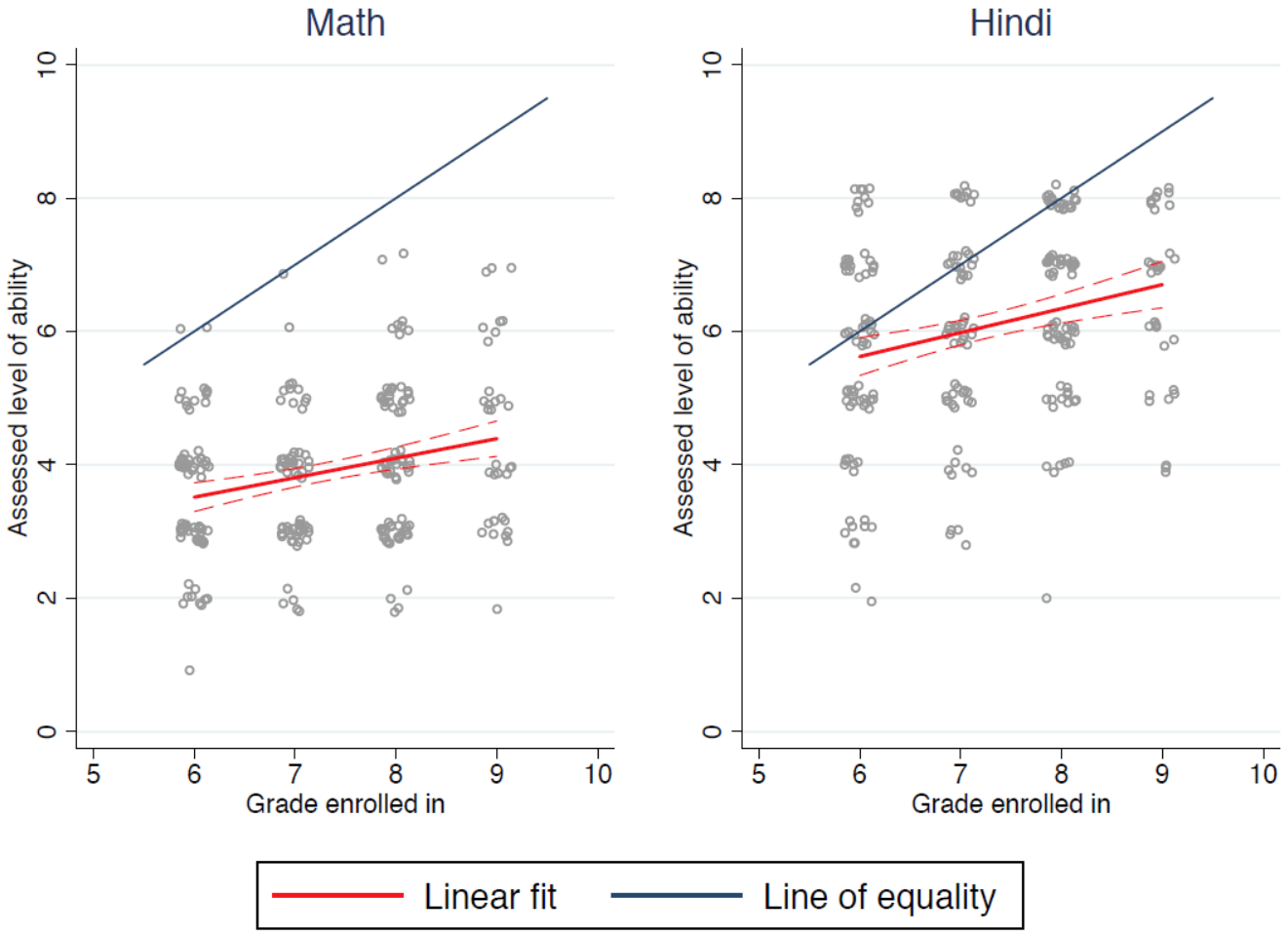
- We present an experimental evaluation of a blended learning program (*Mindspark*) in India
  - Developed by Educational Initiatives (EI), a leading Indian education assessments provider, over a 10-year period
    - Over 45,000 question Item Bank, used by over 400,000 students, administering over a million questions daily
  - Provides individual, dynamically updated, assessment and content
  - Instruction is targeted at children's actual level of achievement, **not the curriculum-mandated level**
  - The software itself is platform-agnostic: deployed in school-based, internet-based, and after-school models
- We evaluate the after-school model (Mindspark centers), which provide supplementary after-school instruction to students six days/week
  - 45 mins individual study using CAL software (Mindspark); 45 mins small group teaching (12-15 students)
  - 619 students, individual level randomization, 4.5 months treatment
  - Treated students received a complete fee waiver to attend Mindspark centers (will present ITT and IV estimates)

Summary of results

- **Business-as-usual learning:**
  - Students in this setting are several grade-levels below their grade-appropriate standard and this gap grows with each grade
  - In the control group, students in the bottom third of the within-grade distribution show zero progress in both math and Hindi on our independently-administered tests
- **Program effects:**
  - The offer of a Mindspark voucher led to large increases of 0.36 SD in math and 0.22 SD in Hindi over the study period (just 4.5 months)
    - IV estimates: 0.59 SD/0.36 SD in math/Hindi in 90 days
  - Large and similar absolute test-score gains for all students
  - Much larger *relative* gains for academically-weaker students
- **Other results:**
  - The CAL system accommodates the wide amount of variation in the classroom (spanning 5-6 grade levels); average student starts 3-4 grade levels below in math and makes rapid progress
  - Grade-level decomposition and school-exam results
  - Highly cost effective (especially productivity per unit of time)

Actual vs. expected learning levels

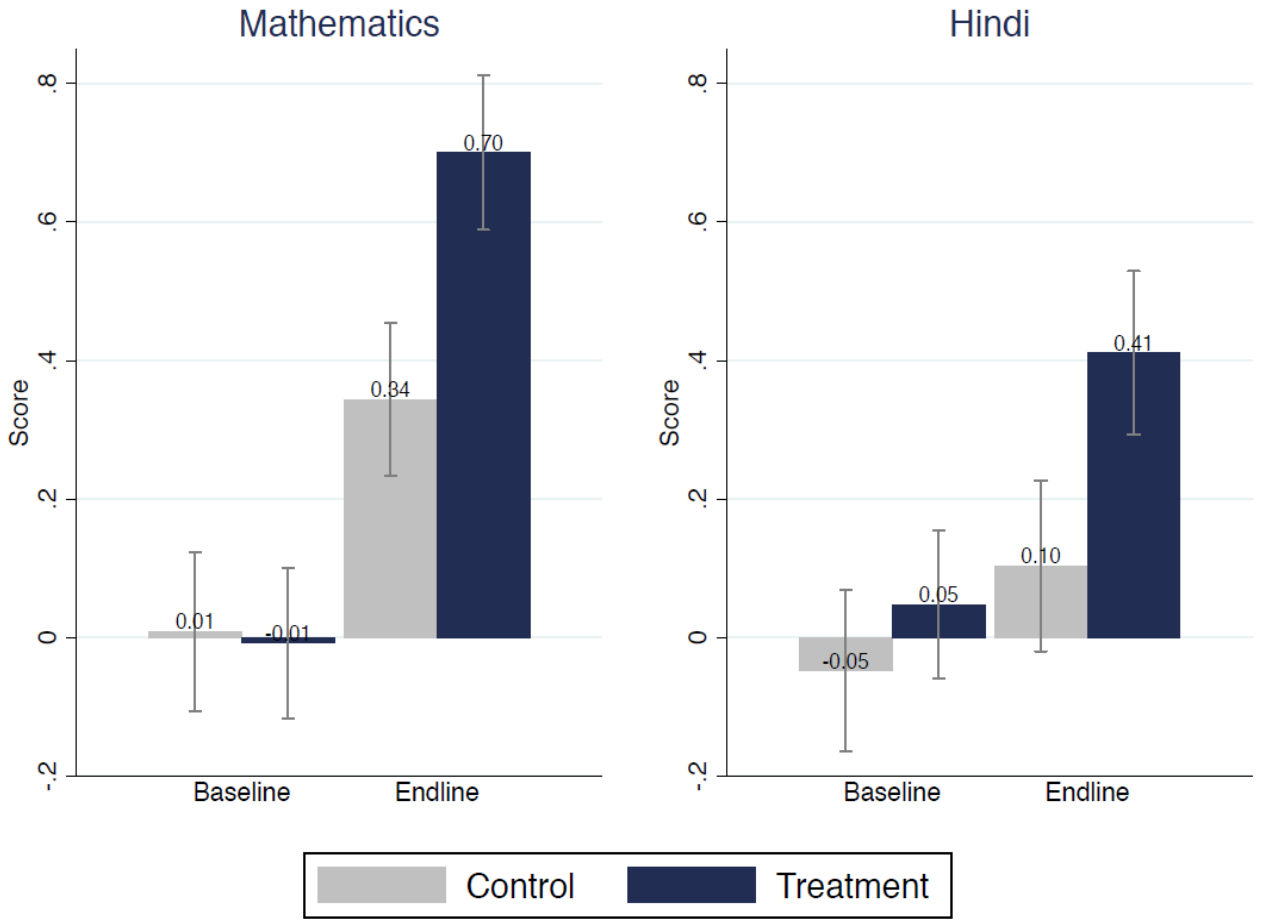
In the treatment group at start of intervention



This figure shows, for treatment group, the actual ability level (determined by the Mindspark CAL program) plotted against the grade they are enrolled in.

The core result

Mean differences in achievement



Test scores were linked within-subject through IRT models, pooling across grades and across baseline and endline, and are normalized to have a mean of zero and a standard deviation of one in the baseline.

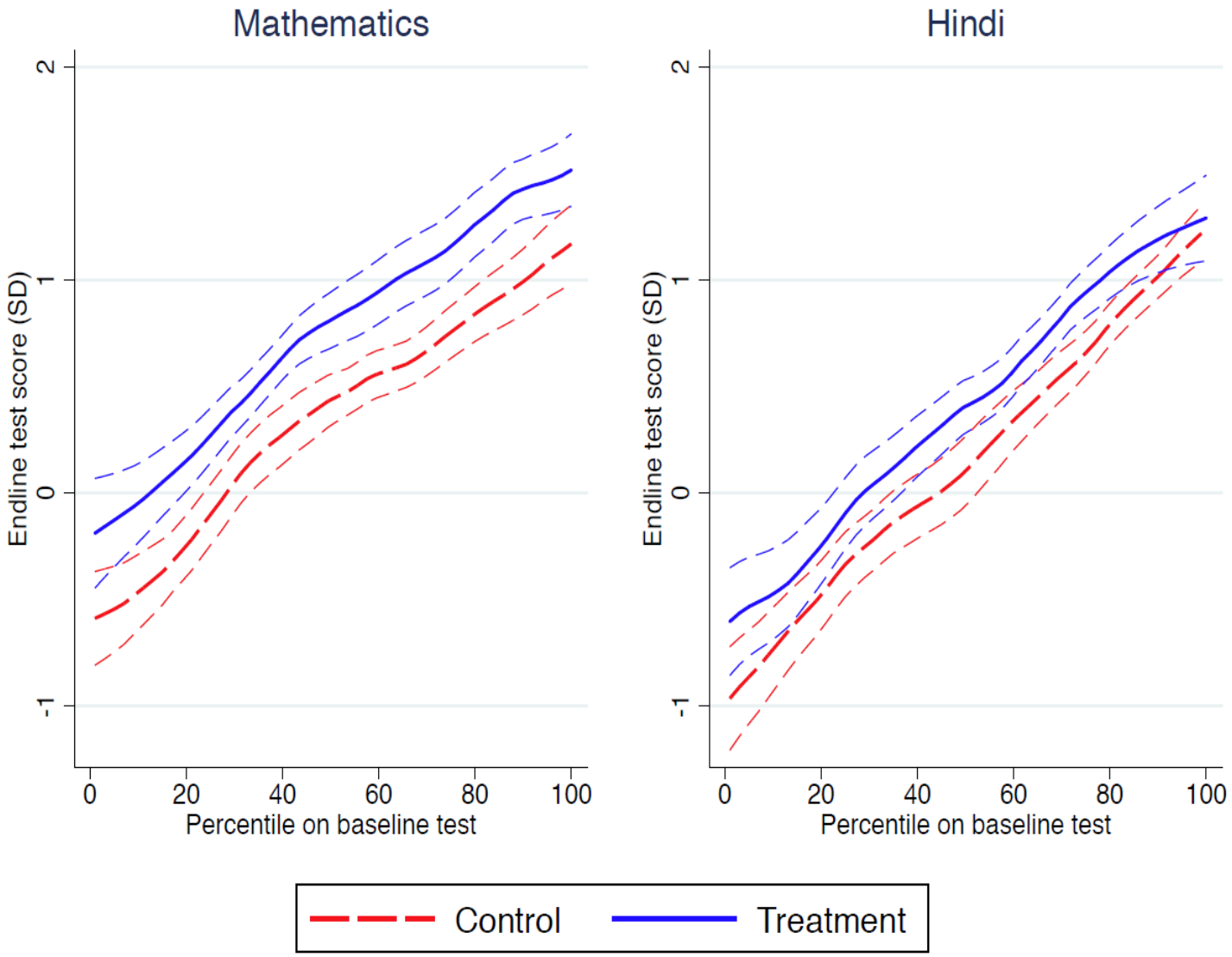
The core result

ITT Results in a regression framework

	(1)	(2)	(3)	(4)
	Dep var: Standardized IRT scores (endline)			
	Math	Hindi	Math	Hindi
Treatment	0.36***	0.22***	0.36***	0.22***
	(0.063)	(0.076)	(0.062)	(0.064)
Baseline score	0.54***	0.67***	0.55***	0.69***
	(0.047)	(0.034)	(0.039)	(0.039)
Constant	0.36***	0.15***	0.36***	0.15***
	(0.031)	(0.038)	(0.043)	(0.045)
Strata fixed effects	Y	Y	N	N
Observations	529	533	529	533
R-squared	0.392	0.451	0.392	0.465

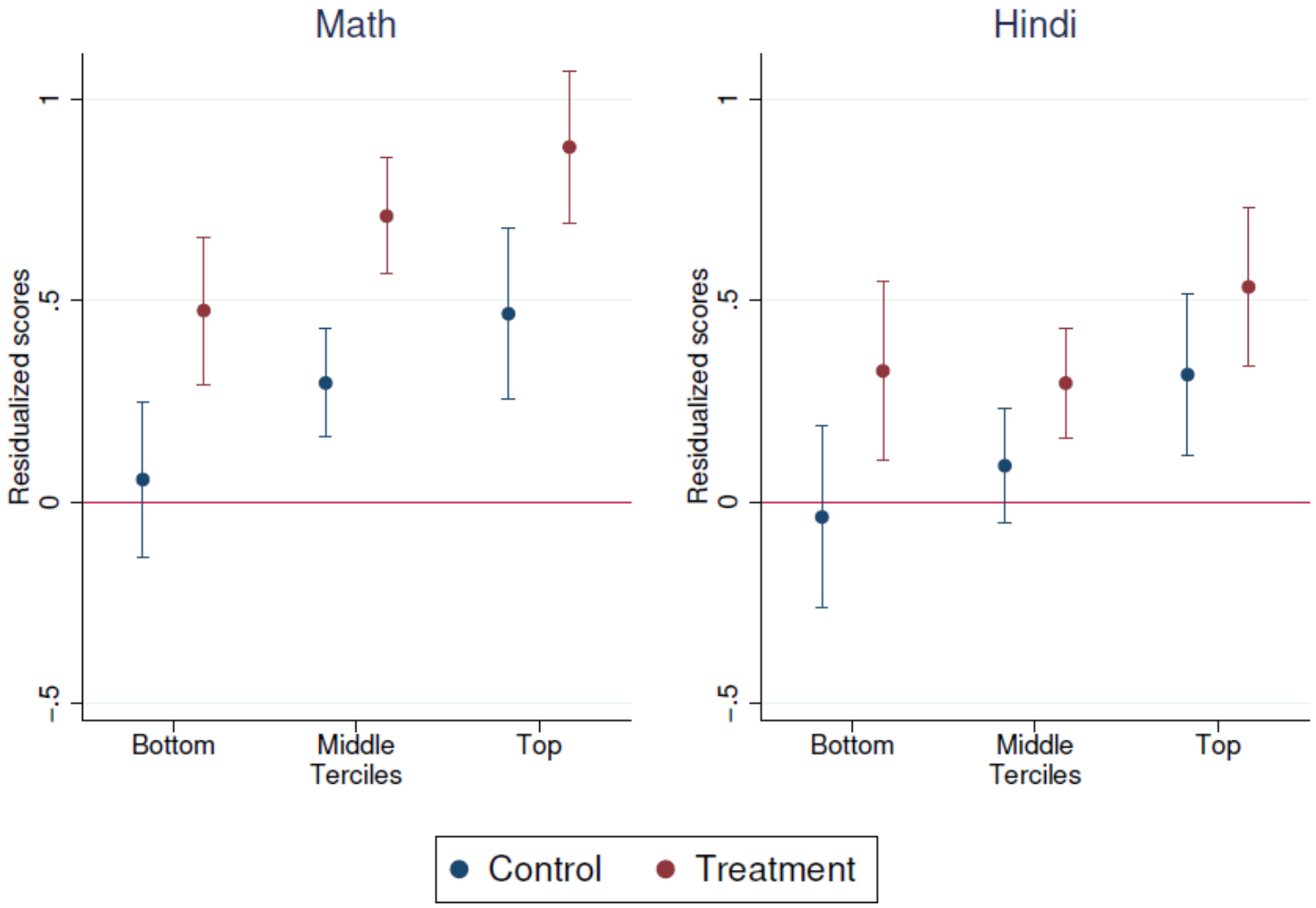
Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 Tests scores are scaled here using Item Response theory models and standardized to have a mean of zero and standard deviation of one in the baseline.

Learning gains across the full distribution



Treatment vs. "business-as-usual" progress

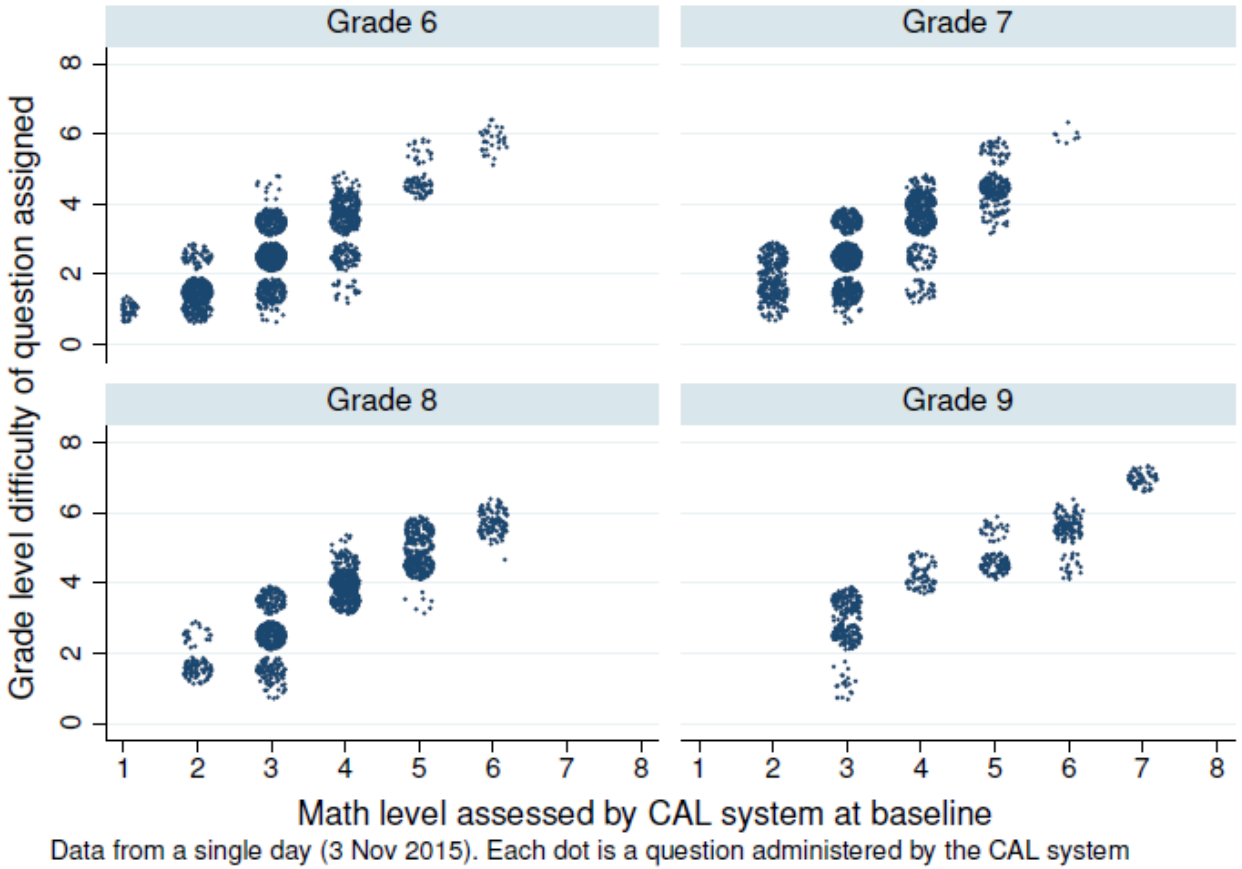
Children in the lowest terciles make *zero progress* in control



Plotted against own quantiles

No single teacher can individualize instruction so finely

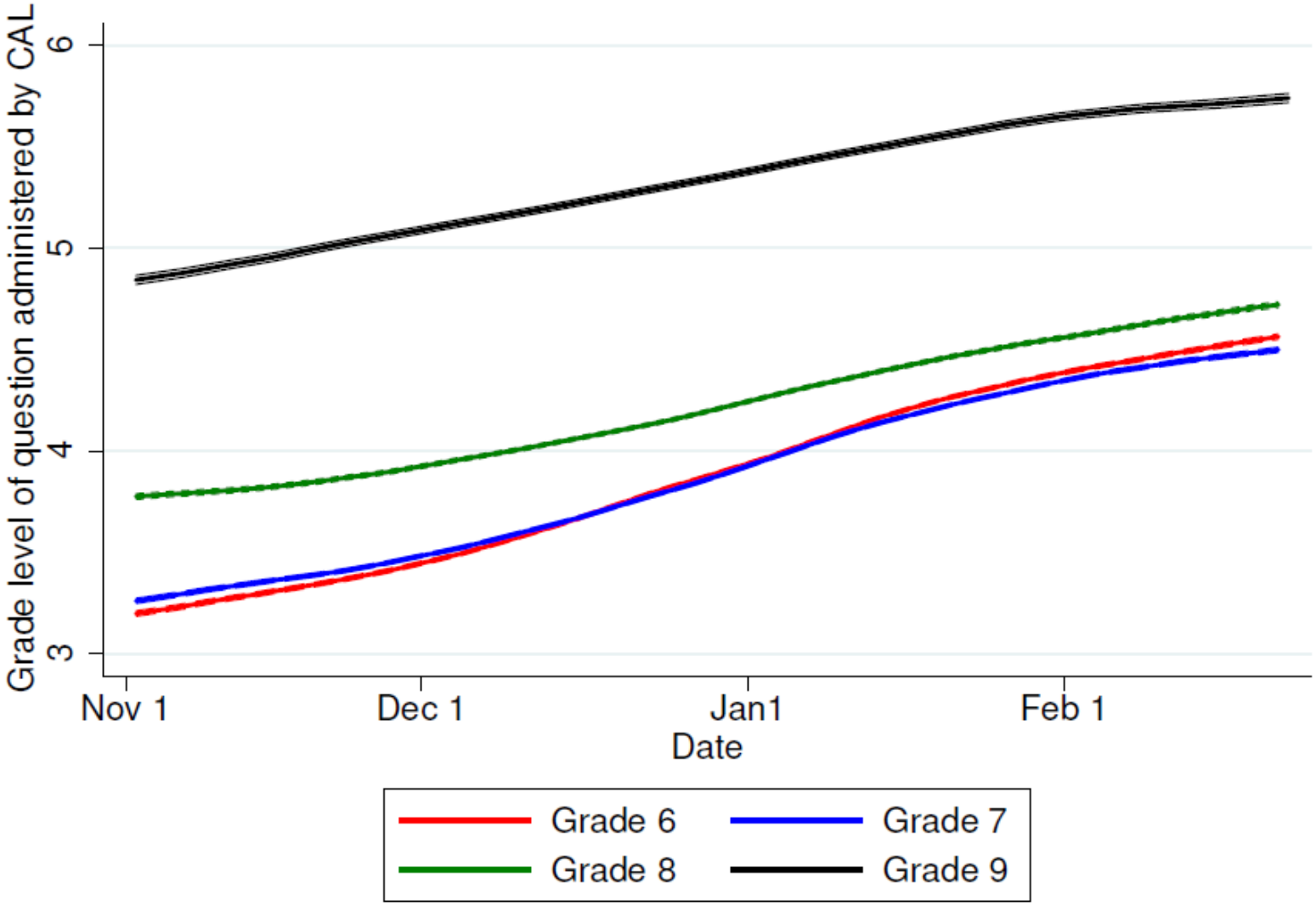
CAL caters to wide range of ability in a single session



This figure shows, for treatment group, the grade level of questions administered by the computer adaptive system to students **in a single day** (3 Nov 2015). The CAL system (a) allows for precise targeting to individual ability levels; (b) can cope with wide variation in ability levels within and across grade levels; (c) can adapt quickly to changes in ability.

Students in all grades learn over the study period

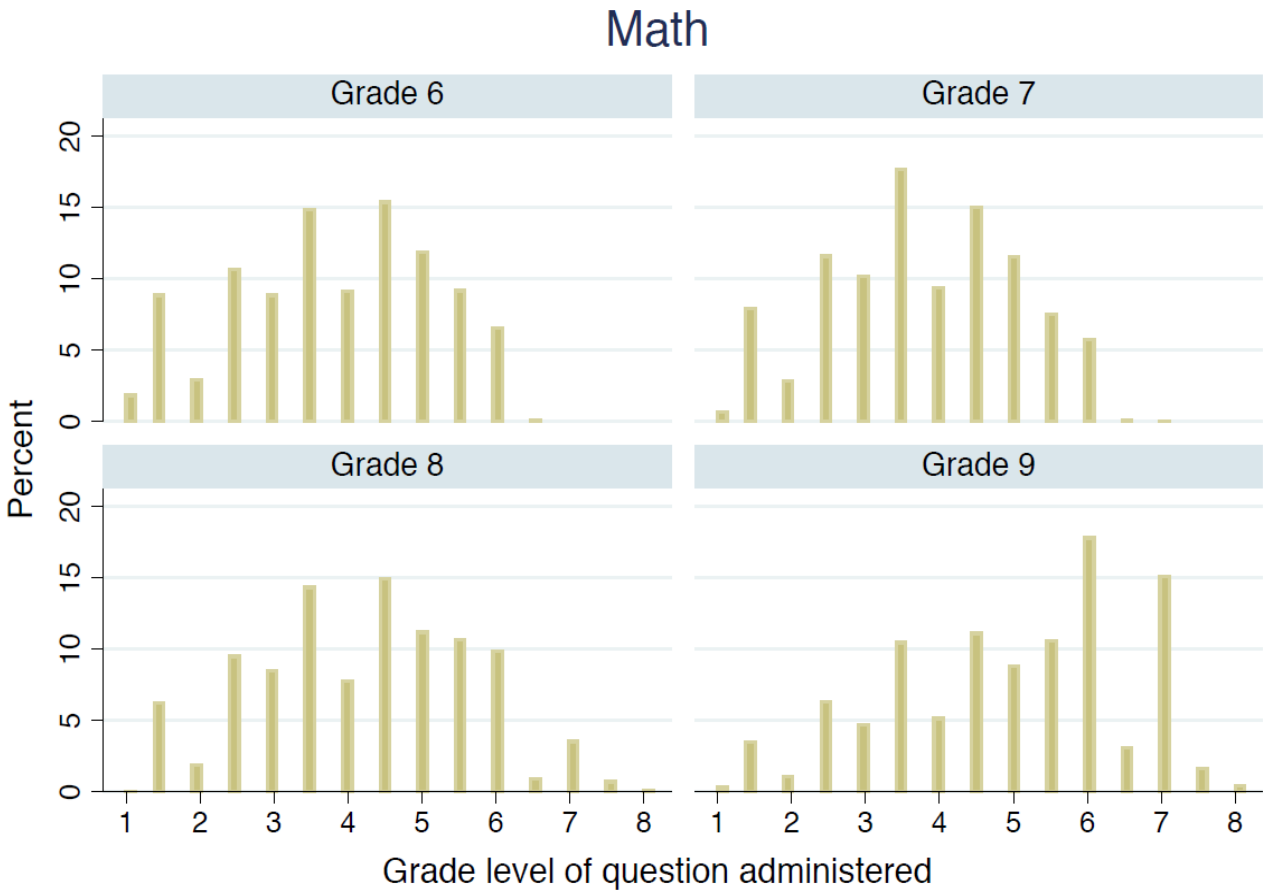
The increase in learning is continuous *and continuously adapted to at individual level*



Plotted separately by actual grade enrolled in

What was the Mindspark CAL system teaching

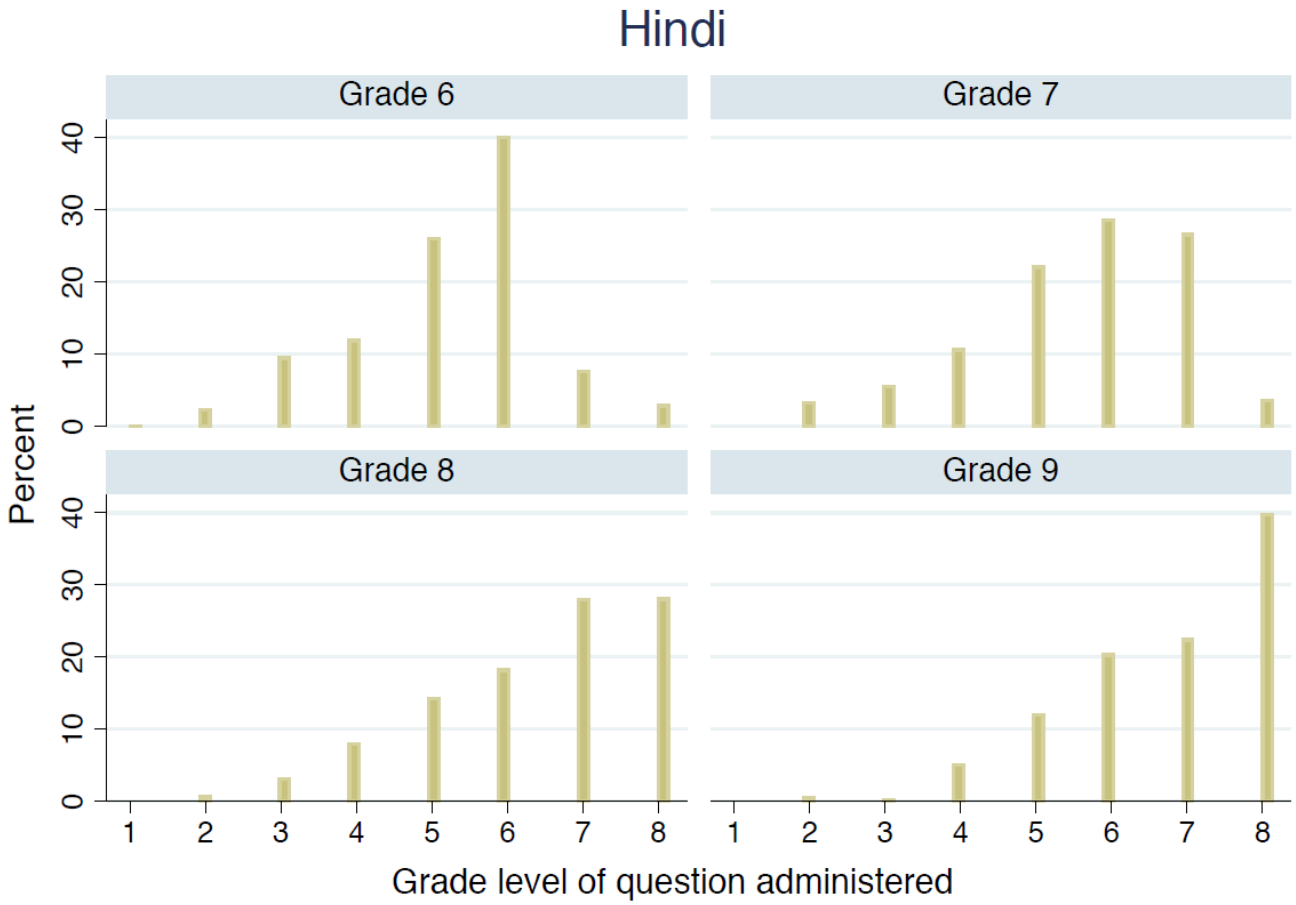
In math, very few questions at grade level



Note: This table aggregates over 380,707 student responses to 13076 individual questions over the program. Students attempted an average of 1510 questions in Math

What was the Mindspark CAL system teaching

In Hindi, many more questions at grade level



Note: This table aggregates over 233,102 student responses to 10177 individual questions over the program. Students attempted an average of 925 questions in Hindi

Treatment effect on items linked to grade levels

	(1)	(2)	(3)	(4)
	Dep var: Proportion of questions answered correctly			
	Math		Hindi	
VARIABLES	At or above grade level	Below grade level	At or above grade level	Below grade level
Treatment	0.0023	0.082***	0.069**	0.051***
	(0.039)	(0.012)	(0.024)	(0.013)
Baseline math score	0.044	0.095***		
	(0.025)	(0.0056)		
Baseline Hindi score			0.11***	0.13***
			(0.016)	(0.0065)
Constant	0.31***	0.49***	0.44***	0.58***
	(0.018)	(0.0058)	(0.012)	(0.0065)
Observations	286	505	287	507
R-squared	0.025	0.341	0.206	0.379

Note: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$ The dependent variable in each regression is the proportion of questions related to the competence that a student answered correctly. All regressions include randomization strata fixed effects.

Treatment effect on school exams

Across all subjects

	(1)	(2)	(3)	(4)	(5)	(6)
			Dep var: Standardized test scores			
VARIABLES	Hindi	Math	Science	Social Sciences	English	Aggregate
Treatment	0.19**	0.058	0.077	0.10	0.080	0.097
	(0.089)	(0.076)	(0.092)	(0.11)	(0.10)	(0.080)
Baseline Hindi score	0.48***		0.28***	0.41***	0.29***	0.33***
	(0.094)		(0.064)	(0.098)	(0.069)	(0.061)
Baseline math score		0.29***	0.10**	0.25***	0.11**	0.16***
		(0.039)	(0.036)	(0.052)	(0.049)	(0.037)
Constant	0.40	0.14	0.88**	0.69	1.11	0.68
	(1.01)	(0.50)	(0.39)	(0.69)	(0.66)	(0.56)
Observations	595	594	593	592	595	595
R-squared	0.188	0.069	0.117	0.173	0.137	0.202

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1 The regression includes strata, grade and school fixed effects.
School exams were held in March 2016 after the completion of the intervention on a common question paper in each grade. Grades are normalized within school*grade to have a mean of zero and a standard deviation of one in the control group. Baseline math and Hindi scores refer to students' scores on the independent assessment administered as part of the study in Sept 2016.

Process Discovery for Scale

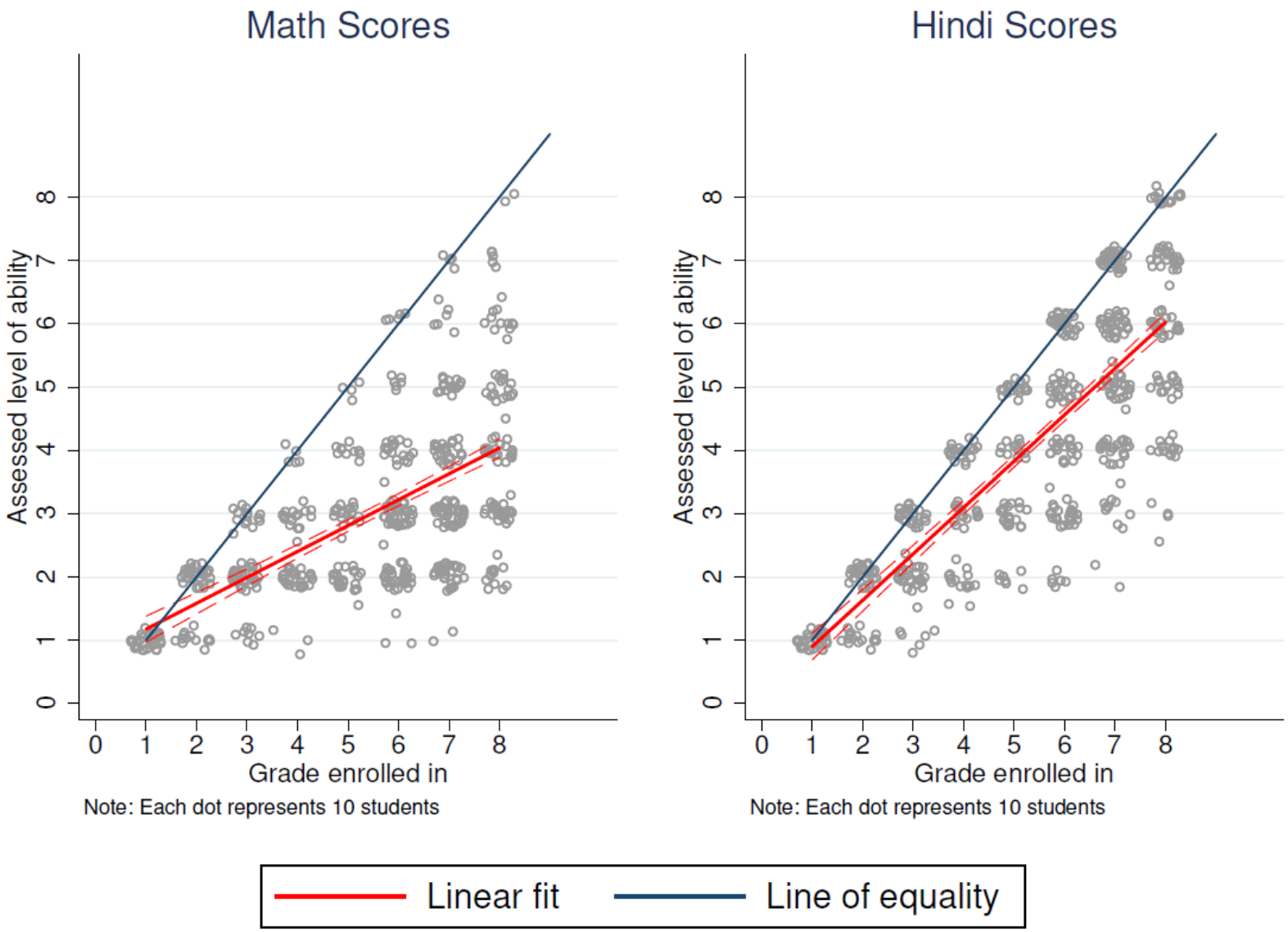
- The Delhi results are promising, but are best considered a "Proof of Concept" that large gains are possible in rapid time frames with a combination of the benefits of CAL and TaRL

- Can we replicate these results in government schools?
  - In Delhi, EI ran all the logistics and the after-school centers
  - Self-selected sample of interested students
  - Supplementing as opposed to substituting (cf with Berry & Mukherjee)

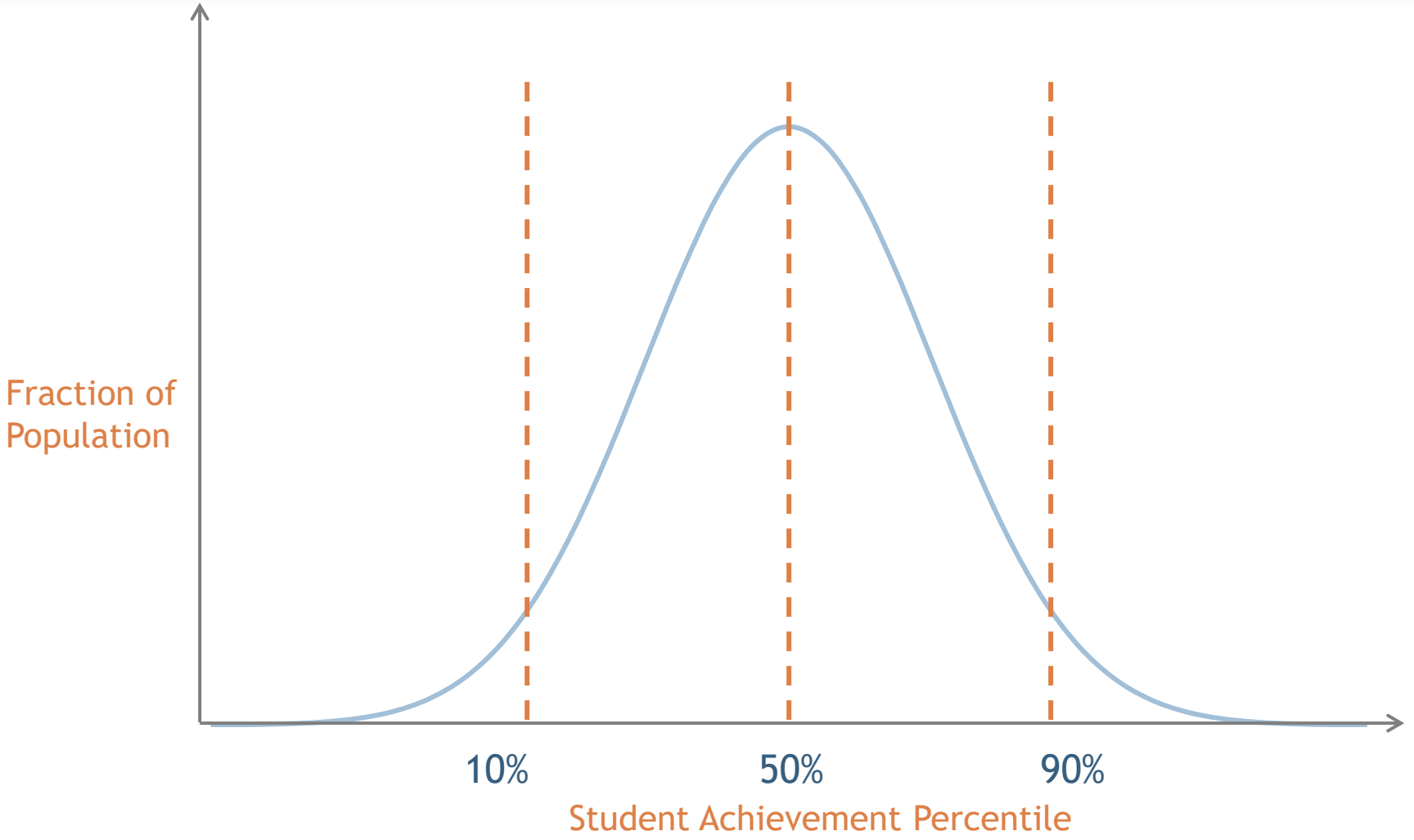
- Currently wrapping up a 3-year study of scale up reaching over 5000 students in 40 schools in state of Rajasthan
  - In-school model with Mindspark periods integrated into the timetable

Mismatch between grade levels and actual achievement

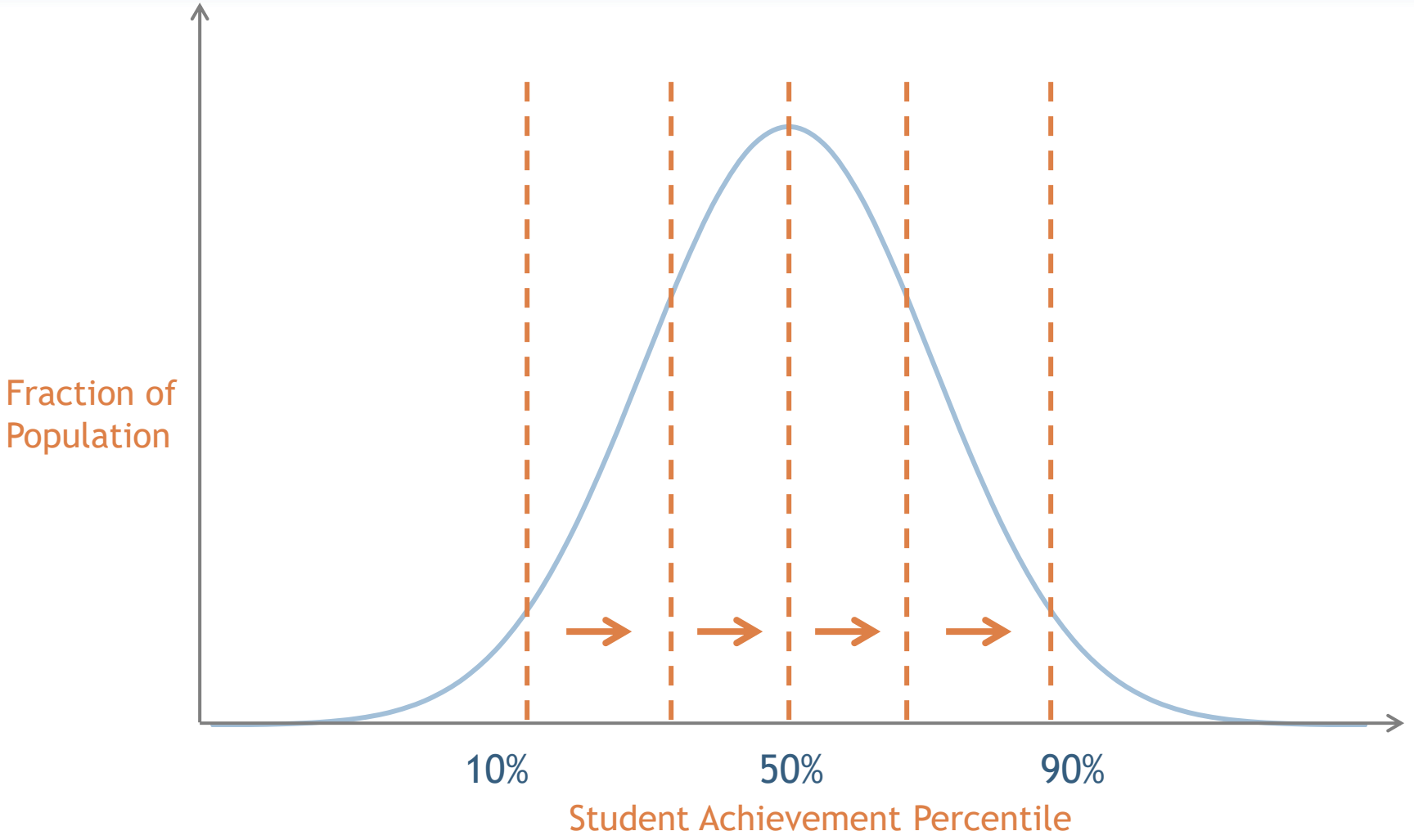
Learning deficits and within-grade dispersion in achievement



Selection Versus Developmental Paradigm in Education



Moving from Selection to Human Capital at all Parts of the Distribution

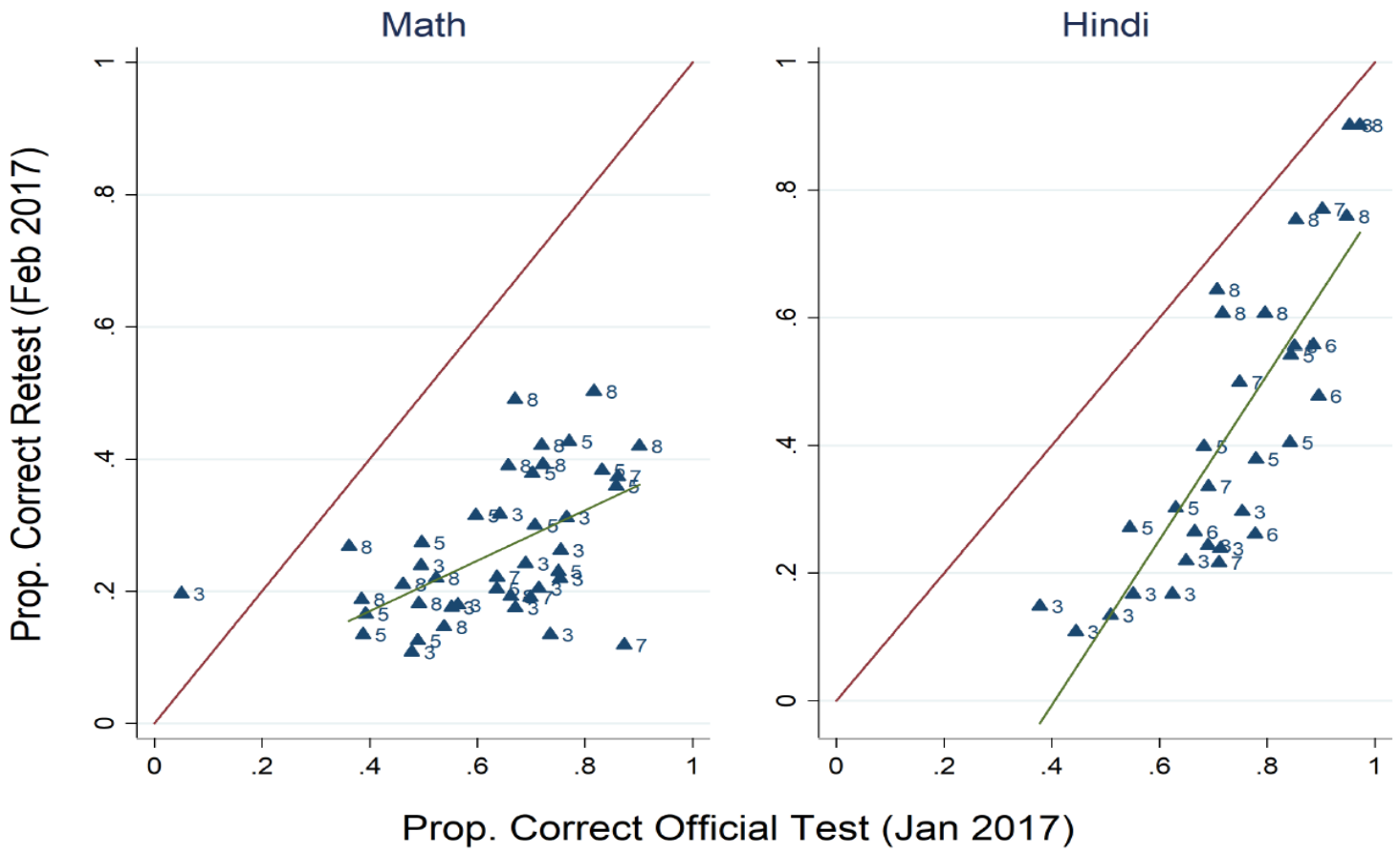


A few recent papers of note

- Bianchi et al (2021)
  - Studies impact of connecting top teachers in China to rural students through broadband internet
  - Finds strong positive impacts on academic achievement; labor market outcomes
  - Finds these effects 7-10 years after the program suggesting long-term impacts
- Navarro-Sola (2021)
  - Studies impact of expanding junior secondary education in Mexico using *telesecunderia* – schools using televised lessons
  - Finds that each year of education increased earnings by 12.5-13.9%
- Beg et al (2021)
  - RCT in Pakistan found that adding expert-led curriculum-based videos to the class raised test scores, but that just giving the same content to students reduced test scores – suggesting key mediating role for teacher
- Derksen et al (2021)
  - Studies impact of Wikipedia access on learning
- Ferman et al (2021)
  - RCT of AI based feedback to students on writing – large positive effects

Technology for Governance? The Challenge of Data Integrity (Singh 2021)

PP Jan 2017 vs. independent assessment in Feb 2017 (Math)

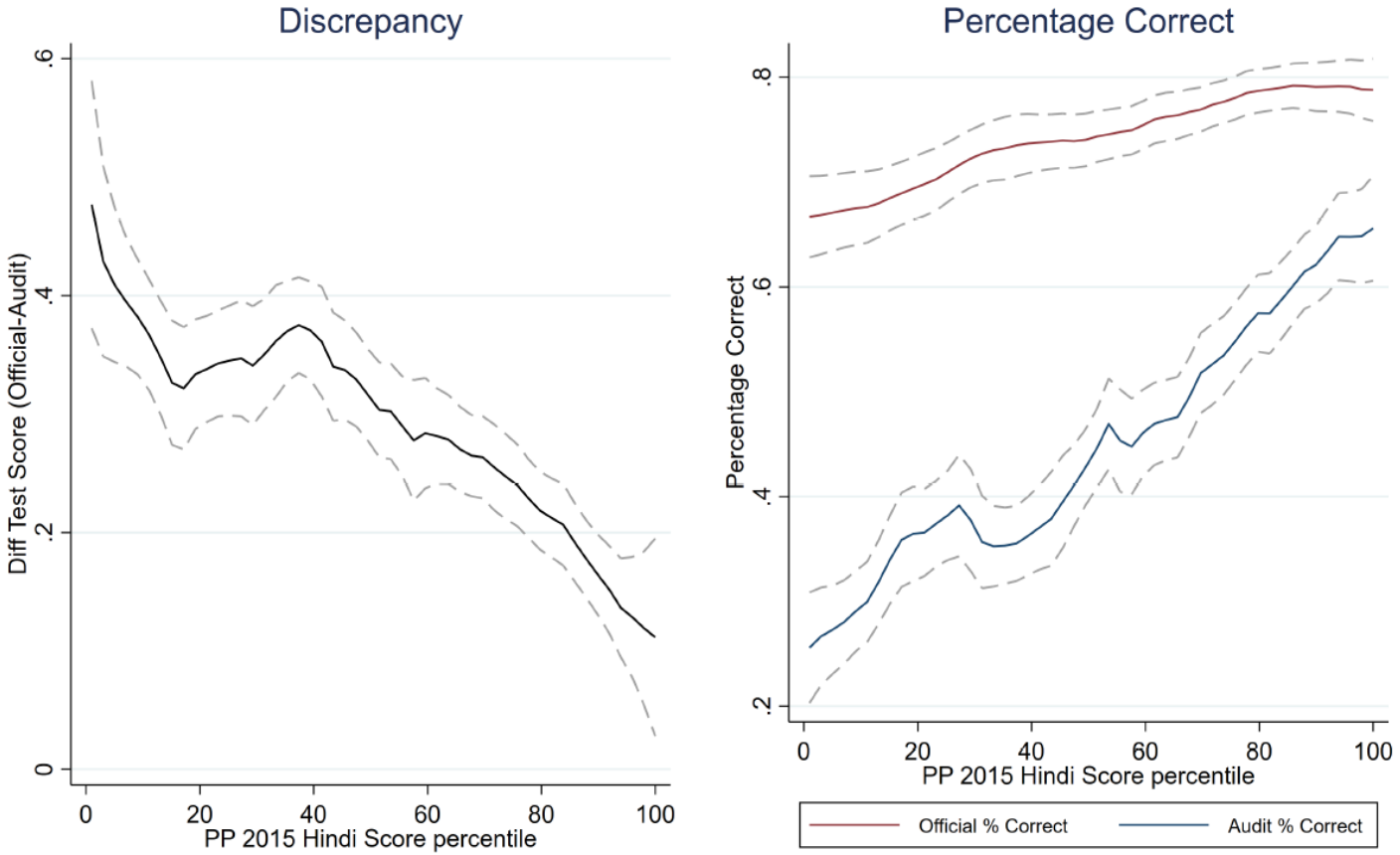


Note:

Each dot shows the proportion of students correctly answering an identical question in PP (Dec 2016) and independent assessment (Feb 2017). Marker labels indicate the grade in which the item was asked.

Overstated at all levels (but more for weaker students)

Discrepancy between official and audit assessments

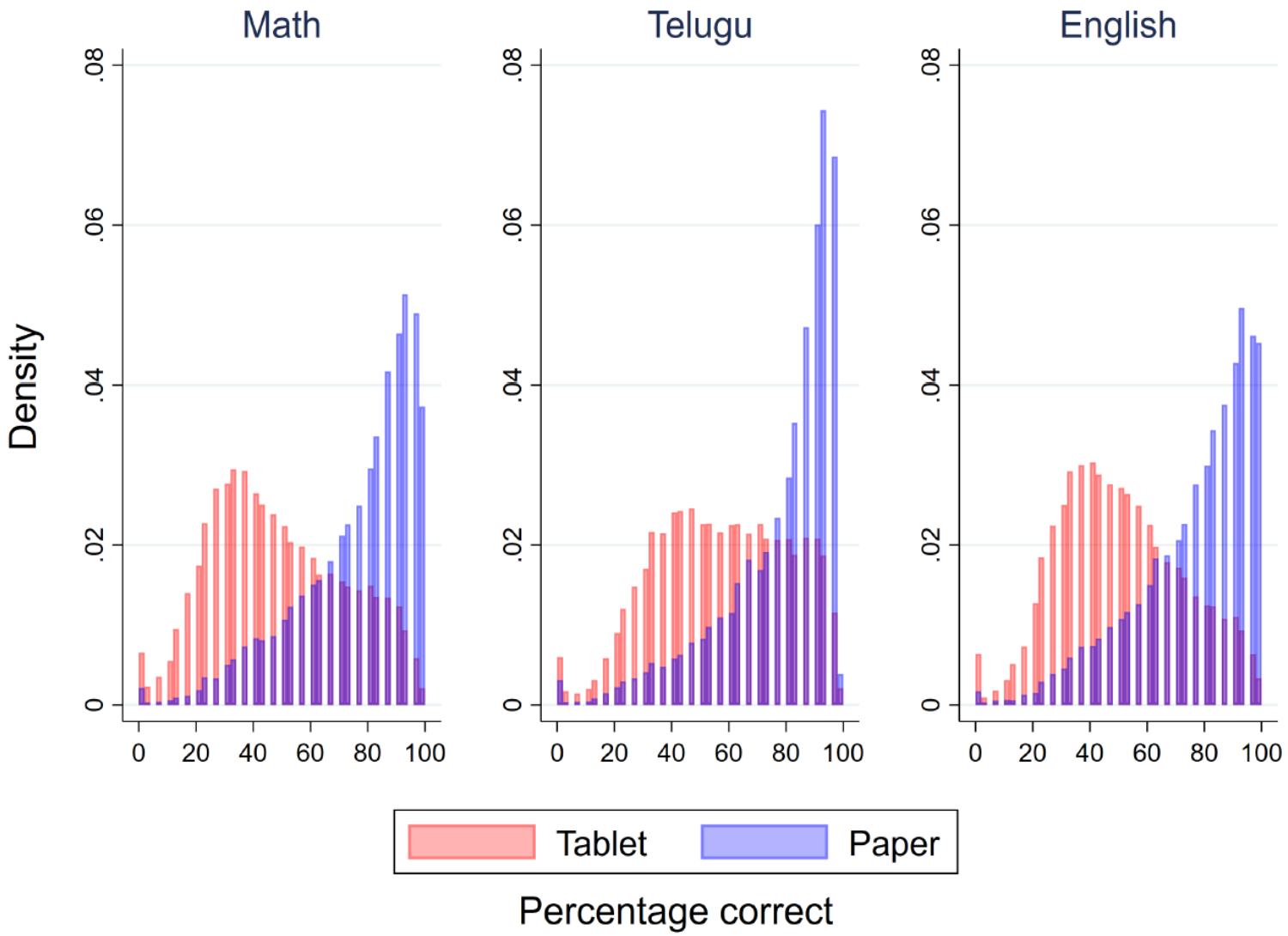


Note:

Discrepancy is defined as the difference between proportion correctly answered in the official assessment and the retest for test items which are common across both assessments. Percentiles are defined over the test score in the previous academic year.

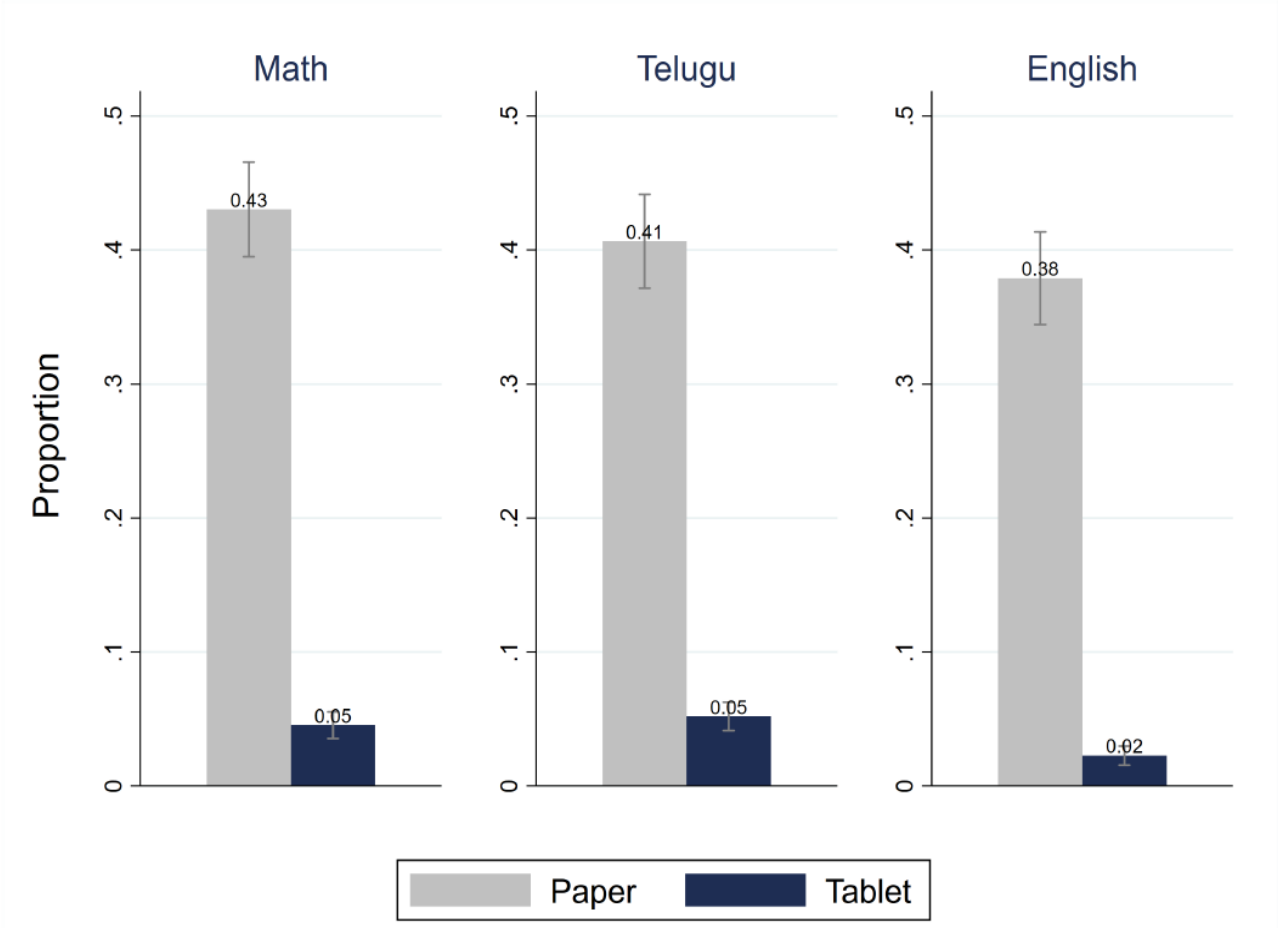
Results

Test score distribution in tablet and paper tests – Student level



The main result

Paper-based tests much more likely to suggest cheating



Note: This figure shows the proportion of schools in the paper and tablet testing arms which are flagged as having potentially cheated based on the procedure in Angrist et al (2017).

Wrap up

- Technology has large potential to improve both pedagogy and governance in school systems
- But requires careful attention to what the binding constraints are and to using technology to alleviate these constraints
- Default focus in most governments is on hardware procurement (and putting pictures of politicians on laptops that are distributed!)
- This is unlikely to have much impact (OLPC, RJ examples)
- Pandemic is likely to have exacerbated inequalities (sometimes low-tech is best)
- Research frontiers include improvements in measurement, sophistication of what the computer does, engaging parents, reorienting teacher training, understanding and increasing student engagement, composite interventions that leverage and test complementarities