Education reforms for system-level change: Evidence from multiple Indian states

Abhijeet Singh

Stockholm School of Economics

RESEP Conference, Stellenbosch University
6th September 2018

Introduction

The Indian context

- Learning levels in India are low and stagnant (or potentially decreasing)
  - Despite a tripling of government spending
  - Unlikely to improve in business-as-usual (Muralidharan 2013).
  - Little evidence that raising inputs is effective
  - Interventions focused on improving governance and pedagogy are much more promising (Glewwe & Muralidharan 2015).

Introduction

The Indian context

- Learning levels in India are low and stagnant (or potentially decreasing)
  - Despite a tripling of government spending
  - Unlikely to improve in business-as-usual (Muralidharan 2013).
  - Little evidence that raising inputs is effective
  - Interventions focused on improving governance and pedagogy are much more promising (Glewwe & Muralidharan 2015).
- Inequality in test scores among the highest in the world
  - in TIMSS, second only to South Africa (Das and Zajonc, 2010)

Introduction

The Indian context

- Learning levels in India are low and stagnant (or potentially decreasing)
  - Despite a tripling of government spending
  - Unlikely to improve in business-as-usual (Muralidharan 2013).
  - Little evidence that raising inputs is effective
  - Interventions focused on improving governance and pedagogy are much more promising (Glewwe & Muralidharan 2015).
- Inequality in test scores among the highest in the world
  - in TIMSS, second only to South Africa (Das and Zajonc, 2010)
- States are responsible for much of frontline education delivery
  - States can differ a lot, both in business-as-usual and interventions

Introduction

The Indian context

- Learning levels in India are low and stagnant (or potentially decreasing)
  - Despite a tripling of government spending
  - Unlikely to improve in business-as-usual (Muralidharan 2013).
  - Little evidence that raising inputs is effective
  - Interventions focused on improving governance and pedagogy are much more promising (Glewwe & Muralidharan 2015).
- Inequality in test scores among the highest in the world
  - in TIMSS, second only to South Africa (Das and Zajonc, 2010)
- States are responsible for much of frontline education delivery
  - States can differ a lot, both in business-as-usual and interventions
- This presentation is largely about our attempt to evaluate policy reforms in multiple states
  - all reforms in the state sector, intended eventually for scale

Agenda

A tale of two education systems

Contrasting India and Vietnam

- India and Vietnam look similar in many respects
  - both middle income countries (~6000-6500 PPP$)
  - both have grown fast in the past 15 years
  - both have seen sharp rise in the average years of education
  - both have seen an increasing share of recent cohorts entering higher education (Sanchez and Singh, 2018)

A tale of two education systems

Contrasting India and Vietnam

- India and Vietnam look similar in many respects
  - both middle income countries (~6000-6500 PPP$)
  - both have grown fast in the past 15 years
  - both have seen sharp rise in the average years of education
  - both have seen an increasing share of recent cohorts entering higher education (Sanchez and Singh, 2018)
- Yet, they differ in a very important dimension – the levels of learning
  - in PISA, Vietnam is the highest performing developing country at roughly the level of UK and Germany (both in 2012 and 2015)
  - India, only entered once for two states, and had the second-worst results in the world (after Kyrgyztan!)

How early do these gaps start?

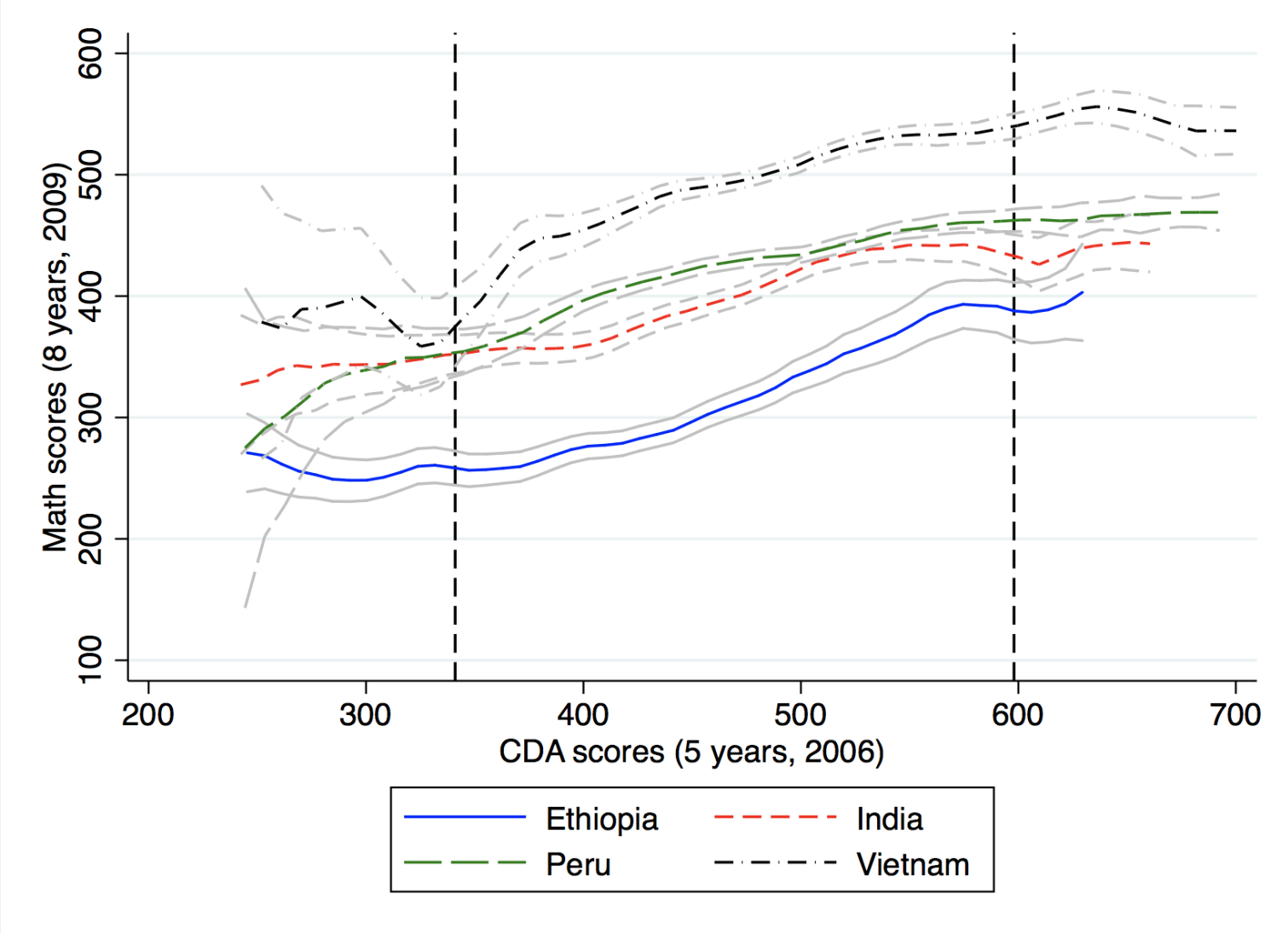
Learning Differences at 5 and 8

		Ethiopia	India	Peru	Vietnam
5 years old (2006)	Mean	412.8	466.7	494.0	500.0
	SD	112.2	100.0	105.0	100.0
	N	1826	1796	1796	1822
8 years old (2009)	Mean	294.4	397.7	427.1	500.0
	SD	119.4	104.8	80.5	100.0
	N	1826	1796	1796	1822

Scores are IRT test scores generated within an age sample, pooling data from all countries, and normalized to have a mean of 500 and an SD of 100 in the Vietnamese sample. Scores are comparable across countries but not across age groups.

How early do these gaps start?

Learning Differences at 5 and 8



Source: Singh (2018)

Does this reflect the productivity of schooling?

	(1)	(2)	(3)	(4)	(5)	(6)
	Dep var: Proportion correct on:					
VARIABLES	Number recognition	Addition	Subtraction	Applied problems	Multiplication	Division
Enrolled grade in 2009	0.094***	0.12***	0.12***	0.15***	0.21***	0.20***
	(0.035)	(0.020)	(0.012)	(0.016)	(0.038)	(0.033)
Grade x Ethiopia	0.059	-0.050**	-0.081***	-0.060***	-0.21***	-0.20***
	(0.037)	(0.023)	(0.014)	(0.018)	(0.038)	(0.033)
Grade x India	-0.034	-0.019	-0.047***	-0.081***	-0.15***	-0.15***
	(0.036)	(0.020)	(0.013)	(0.018)	(0.039)	(0.034)
Grade x Peru	-0.0062	0.025	-0.013	-0.0075	-0.11**	-0.11***
	(0.036)	(0.024)	(0.016)	(0.019)	(0.044)	(0.038)
Age in months	-0.00083	0.0011	0.0021***	0.0036***	0.0056***	0.0047***
	(0.0012)	(0.00093)	(0.00062)	(0.00092)	(0.0013)	(0.0011)
Test score (2006)	0.00020***	0.00033***	0.00031***	0.00036***	0.00020***	0.00016***
	(0.000028)	(0.000038)	(0.000032)	(0.000042)	(0.000032)	(0.000029)
Constant	0.50***	-0.16*	-0.33***	-0.38***	-0.78***	-0.67***
	(0.10)	(0.090)	(0.068)	(0.090)	(0.13)	(0.11)
Observations	7,130	7,130	7,130	7,130	7,130	7,130
R-squared	0.201	0.236	0.205	0.201	0.144	0.137
	Country-specific mean of the dependent variable					
Ethiopia	0.58	0.20	0.11	0.37	0.01	0.01
India	0.89	0.45	0.27	0.56	0.12	0.09
Peru	0.93	0.67	0.37	0.54	0.09	0.07
Vietnam	0.96	0.73	0.65	0.71	0.37	0.36

The challenge for the Indian education system

- The challenge for the Indian education system is of improving learning quality
  - Low level that has been stagnant for over 10 years
  - Over 50% of Grade 5 students can't read at Grade 2 level!
- If we could improve the per-year productivity of the Indian system to Vietnamese levels, we would close 90% of the gap with Vietnam at 8
  - this is despite keeping initial levels of achievement unchanged!
- Change of this magnitude is hard, and especially so at scale
  - but various state governments are trying
- The rest of this presentation is about our attempts to evaluate various policy reforms

Agenda

Why evidence on schooling reforms is necessary

- Existing reforms may be **ineffective**
  - but governments will continue to spend money/time on them because they believe the reforms work.

Why evidence on schooling reforms is necessary

- Existing reforms may be **ineffective**
  - but governments will continue to spend money/time on them because they believe the reforms work.
- Reforms may be **partly effective**
  - but there is no feedback to improve them.

Why evidence on schooling reforms is necessary

- Existing reforms may be **ineffective**
  - but governments will continue to spend money/time on them because they believe the reforms work.
- Reforms may be **partly effective**
  - but there is no feedback to improve them.
- Reforms may be **effective** but
  - they end (due to political opposition or turnover)
  - or not replicated in other settings (due to lack of awareness).

Why evidence on schooling reforms is necessary

- Existing reforms may be **ineffective**
  - but governments will continue to spend money/time on them because they believe the reforms work.
- Reforms may be **partly effective**
  - but there is no feedback to improve them.
- Reforms may be **effective** but
  - they end (due to political opposition or turnover)
  - or not replicated in other settings (due to lack of awareness).
- This presentation will focus on results from different projects
  - goal not to push for specific interventions
  - more for an illustration of what some states in India are trying and how evaluations can feed into this

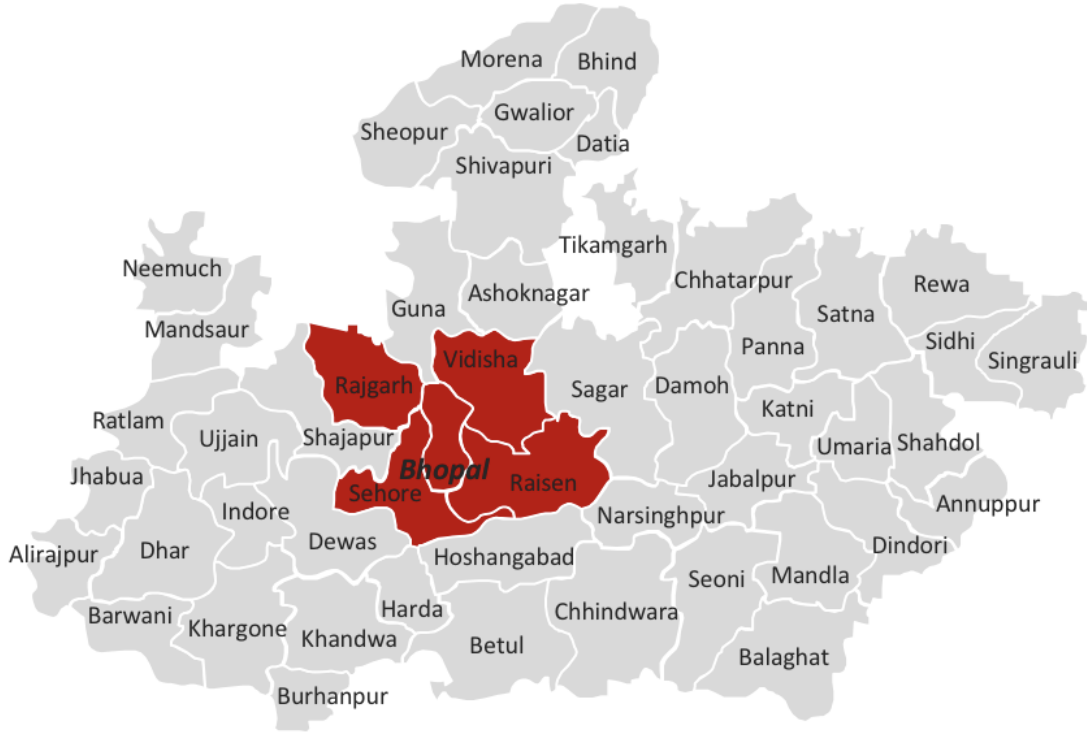
Public Sector Reforms at Scale

- The first two reforms were studied in Madhya Pradesh state and focus on governance and accountability
  - fifth largest state in India by population
  - 72.6 million people in 2011, 72% rural
  - More deprived than the all-India average

Public Sector Reforms at Scale

- The first two reforms were studied in Madhya Pradesh state and focus on governance and accountability
  - fifth largest state in India by population
  - 72.6 million people in 2011, 72% rural
  - More deprived than the all-India average
- The public education system illustrates national problems, with greater severity:
  - Learning levels are low (e.g. 31% of Grade 5 students can read a Grade 2 text)
  - Teacher absence is high, possibly getting worse (26% in 2010, 8 pp worse than 2003)
  - Student attendance is low, and getting worse (68% in 2010, 55% in 2016)
  - Schools are small, and getting smaller (18% of schools with less than 60 students in 2010, 40% in 2016)

Districts in the evaluations



Reform 1: Test-based accountability at scale

- The Govt of MP instituted annual standardized state assessments administered to all students in Grades 1–8 from 2011
- **Objectives** (as per official circulars):

Reform 1: Test-based accountability at scale

- The Govt of MP instituted annual standardized state assessments administered to all students in Grades 1–8 from 2011
- **Objectives** (as per official circulars):
  1. Diagnosis: **Understanding the true situation of quality** of education and **tracking progress** at regular intervals
  2. Signalling: To **demonstrate the commitment and priority** of the government towards children's educational achievements and to sensitize society
  3. Action: To **set up programs and strategies** to improve the academic achievements of the children of the state
  4. Remediation: To **take therapeutic / diagnostic steps** for the children not able to achieve the required skills
  5. Sensitization: To **educate teachers, education administration, public representatives and society** towards the educational achievements of children
  6. Levers: To know the correct position of the school operations and facilities and make necessary corrections.

The Pratibha Parv assessments

Content and administration

- **Content:** Grade-appropriate test questions closely linked to the state curricula
  - in lower grades, mix of individual-oral questions and written
  - in higher grades, exclusively written
  - mix of MCQs and open-ended responses
  - most questions focused on knowledge and recall of facts

The Pratibha Parv assessments

Content and administration

- **Content:** Grade-appropriate test questions closely linked to the state curricula
  - in lower grades, mix of individual-oral questions and written
  - in higher grades, exclusively written
  - mix of MCQs and open-ended responses
  - most questions focused on knowledge and recall of facts
- **Administration:**
  - standardized test paper designed at state HQ, sent to schools
  - tests in school over two days, graded by teachers
    - aggregated scores available digitally, student-wise scores on paper in schools
  - several anti-cheating measures developed over time
    - "flying squads" of external inspectors
    - for Grades 5 and 8, test papers sent to another school for grading
    - multiple sets of question papers to avoid copying by students
  - one day of extra-curricular activities, communicating with parents/officials

The Pratibha Parv assessments

Letter grade thresholds

- A key objective of the assessments is to enable therapeutic/diagnostic steps that improve the status of learning
  - a focus on both **diagnosing who is below-par** and on **directing effort and attention** to them

The Pratibha Parv assessments

Letter grade thresholds

- A key objective of the assessments is to enable therapeutic/diagnostic steps that improve the status of learning
  - a focus on both **diagnosing who is below-par** and on **directing effort and attention** to them
- The diagnosis classifies student/school scores into discrete letter-grades
  - A Grade: 76 and above
  - B Grade: 61-75
  - C Grade: 46-60
  - D Grade: 34-45
  - E Grade: 0-33

The Pratibha Parv assessments

Letter grade thresholds

- A key objective of the assessments is to enable therapeutic/diagnostic steps that improve the status of learning
  - a focus on both **diagnosing who is below-par** and on **directing effort and attention** to them
- The diagnosis classifies student/school scores into discrete letter-grades
  - A Grade: 76 and above
  - B Grade: 61-75
  - C Grade: 46-60
  - D Grade: 34-45
  - E Grade: 0-33
- These grades are applied both to schools and to students
  - for students, D & E is supposed to target remedial instruction
  - for schools, D & E is supposed to target extra focus on school improvement (e.g. Shaala Siddhi)

Important unknowns

- **How reliable are data from such assessments?**
  - Test score manipulation documented in OECD settings (Jacob and Levitt 2003, Angrist and Battistin 2017, Borcan et al. 2017)
  - Governance and monitoring plausibly weaker in LMICs

Important unknowns

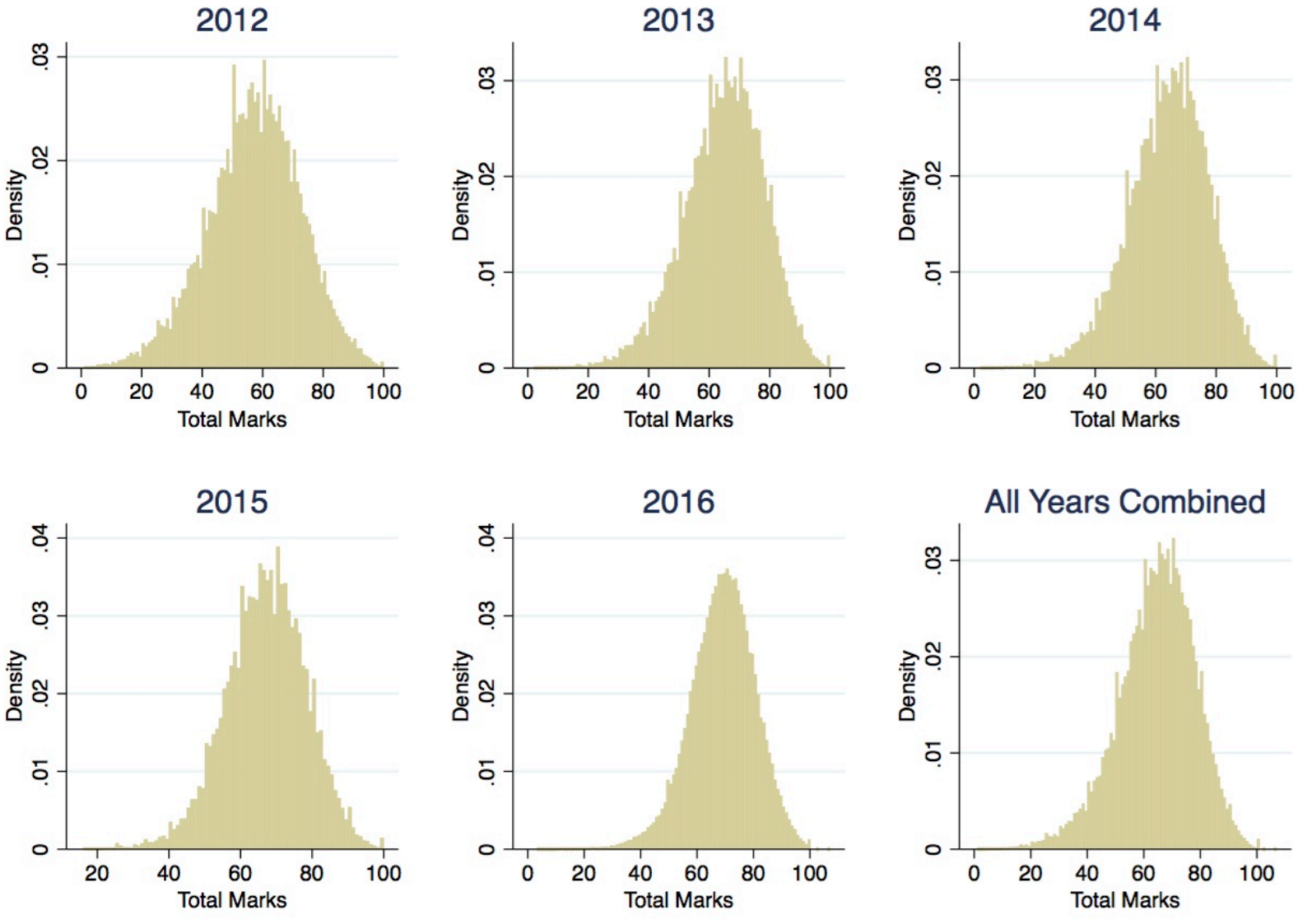
- **How reliable are data from such assessments?**
  - Test score manipulation documented in OECD settings (Jacob and Levitt 2003, Angrist and Battistin 2017, Borcan et al. 2017)
  - Governance and monitoring plausibly weaker in LMICs
- **How effective are current uses of this data?**
  - for formative assessments?
  - for instituting accountability mechanisms?
  - for measuring progress over time?

Important unknowns

- **How reliable are data from such assessments?**
  - Test score manipulation documented in OECD settings (Jacob and Levitt 2003, Angrist and Battistin 2017, Borcan et al. 2017)
  - Governance and monitoring plausibly weaker in LMICs
- **How effective are current uses of this data?**
  - for formative assessments?
  - for instituting accountability mechanisms?
  - for measuring progress over time?
- **What can these data be used for?**
  - No test is perfect, but many can be useful
  - Even with weaknesses in reliability and administration, it might be possible to use these assessments effectively

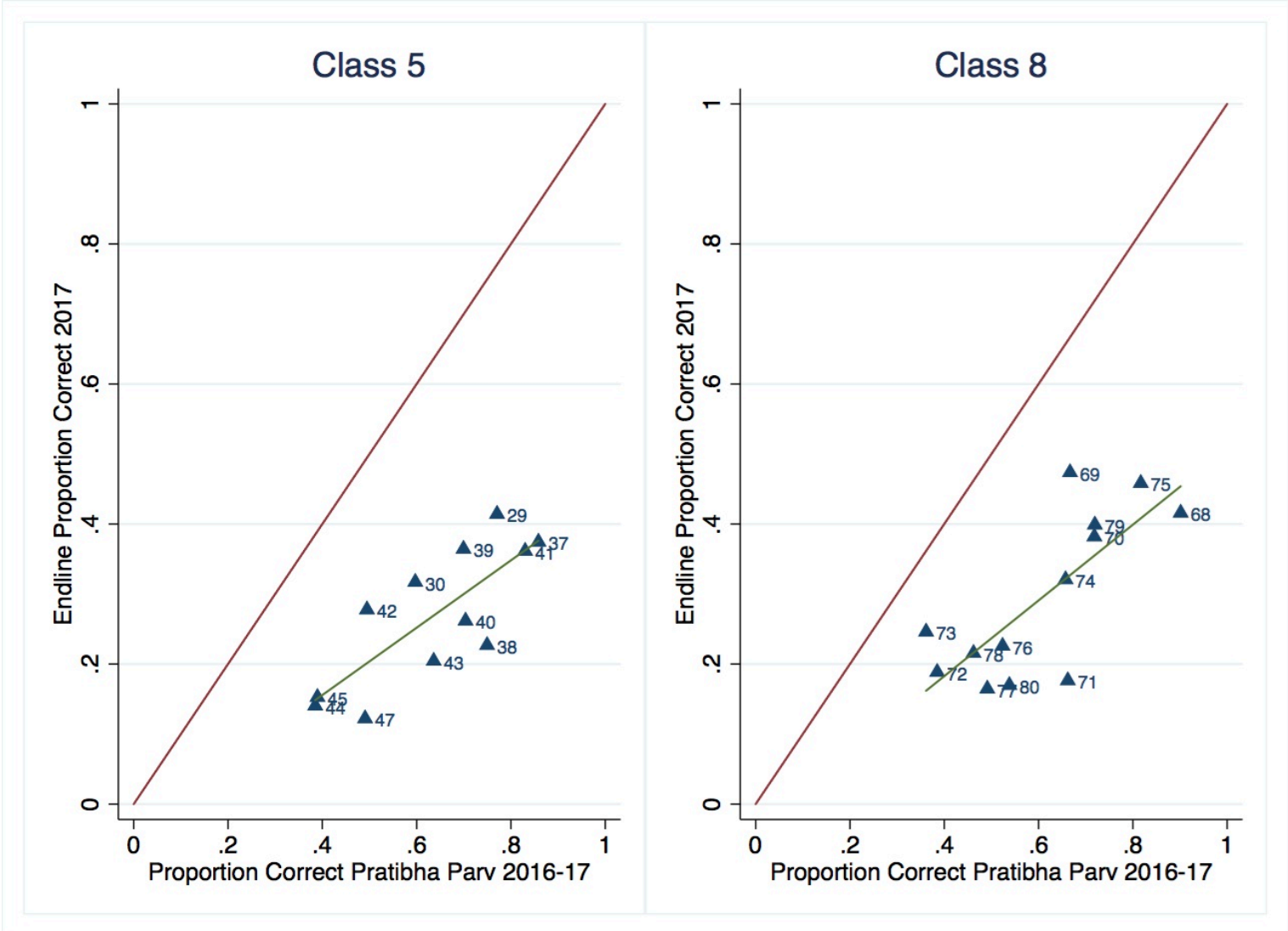
Pratibha Parv marks are mostly well-distributed

Although getting tighter, more skewed over time



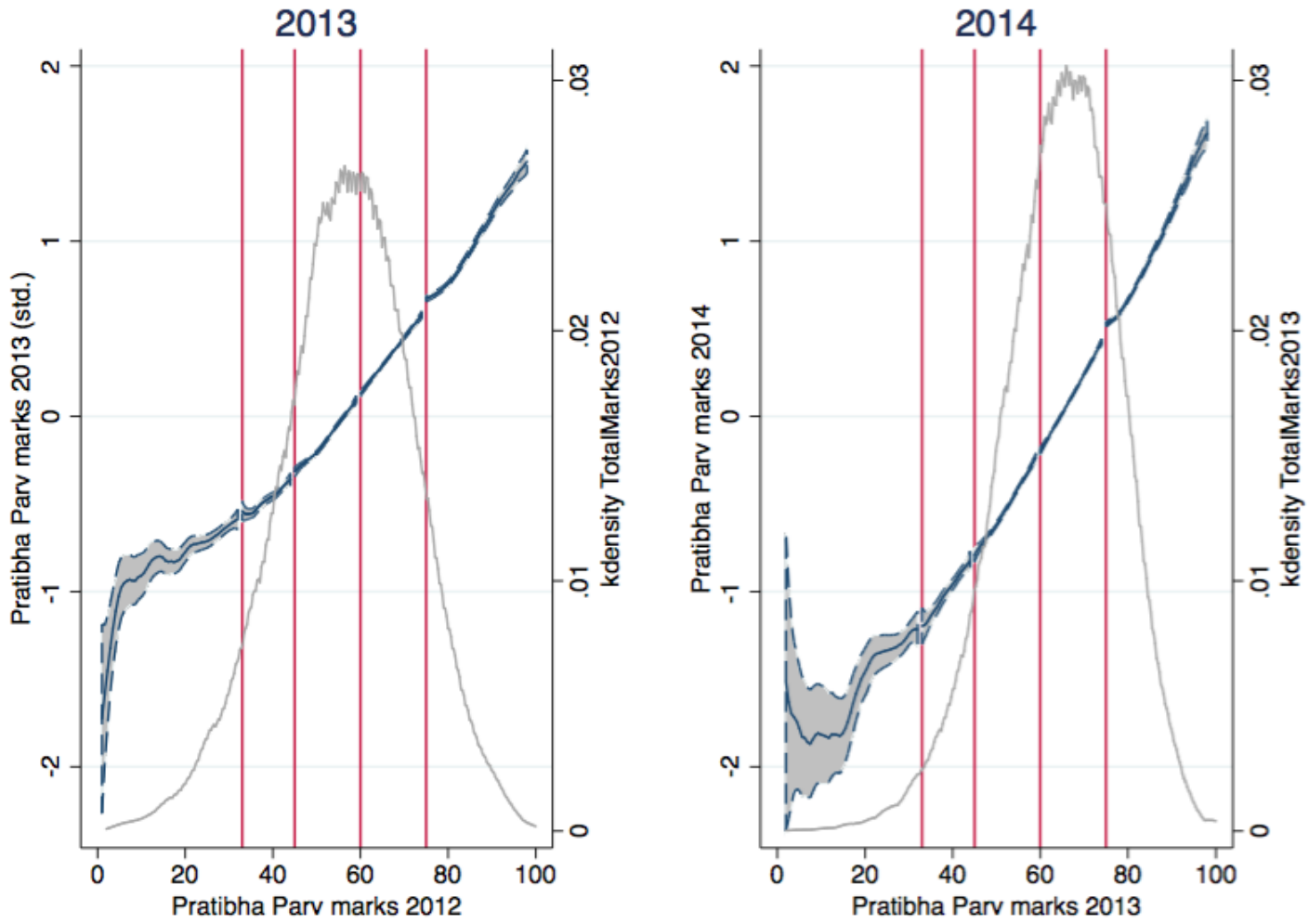
The level of student achievement is **overstated** in PP

PP 2016 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).

No evidence that letter-grades matter for future scores



Note: Lines are local linear smoothed lines plotted separately within each letter grade. We see no evidence of a break at any of the thresholds.

Reform 2: Comprehensive School Evaluations

The MP School Quality Assessment intervention (Muralidharan and Singh, 2018)

- The program (MPSQA, Shaala Gunwatta) is modelled after global "best practices" in school accountability
  - Based on strong political and bureaucratic support within the state
  - With substantial technical inputs from Ark, funded by DFID
  - Modelled on the UK system for school inspections and accountability (Ofsted) with substantial adaptation to context over 2 years

Reform 2: Comprehensive School Evaluations

The MP School Quality Assessment intervention (Muralidharan and Singh, 2018)

- The program (MPSQA, Shaala Gunwatta) is modelled after global "best practices" in school accountability
  - Based on strong political and bureaucratic support within the state
  - With substantial technical inputs from Ark, funded by DFID
  - Modelled on the UK system for school inspections and accountability (Ofsted) with substantial adaptation to context over 2 years
- The goal is to substantially improve school management
- Draws on prevailing understanding that in management theory (Gibbons and Henderson, 2011) that low-performing organizations may stay unproductive because they may not:

Reform 2: Comprehensive School Evaluations

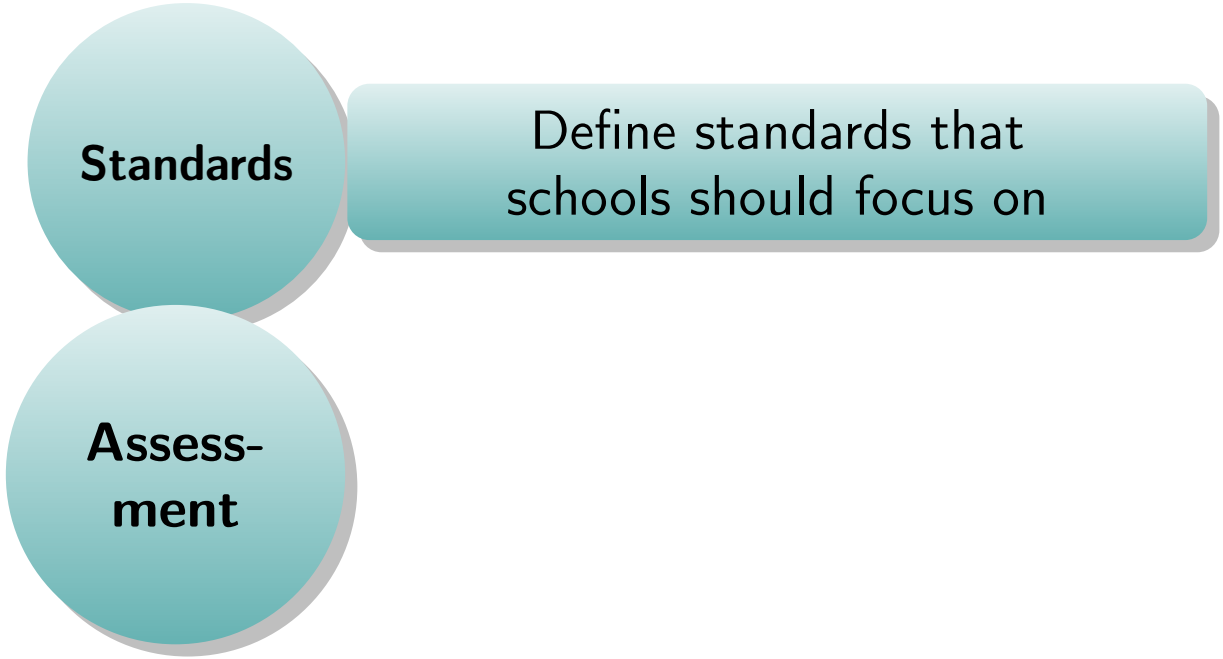
The MP School Quality Assessment intervention (Muralidharan and Singh, 2018)

- The program (MPSQA, Shaala Gunwatta) is modelled after global "best practices" in school accountability
  - Based on strong political and bureaucratic support within the state
  - With substantial technical inputs from Ark, funded by DFID
  - Modelled on the UK system for school inspections and accountability (Ofsted) with substantial adaptation to context over 2 years
- The goal is to substantially improve school management
- Draws on prevailing understanding that in management theory (Gibbons and Henderson, 2011) that low-performing organizations may stay unproductive because they may not:
  - (a) Know they are performing poorly
  - (b) Know what they need to do to improve
  - (c) Be motivated or held accountable for improvements
  - (d) Or they are affected by external factors beyond their control
- **The program tries to address each of these constraints.**

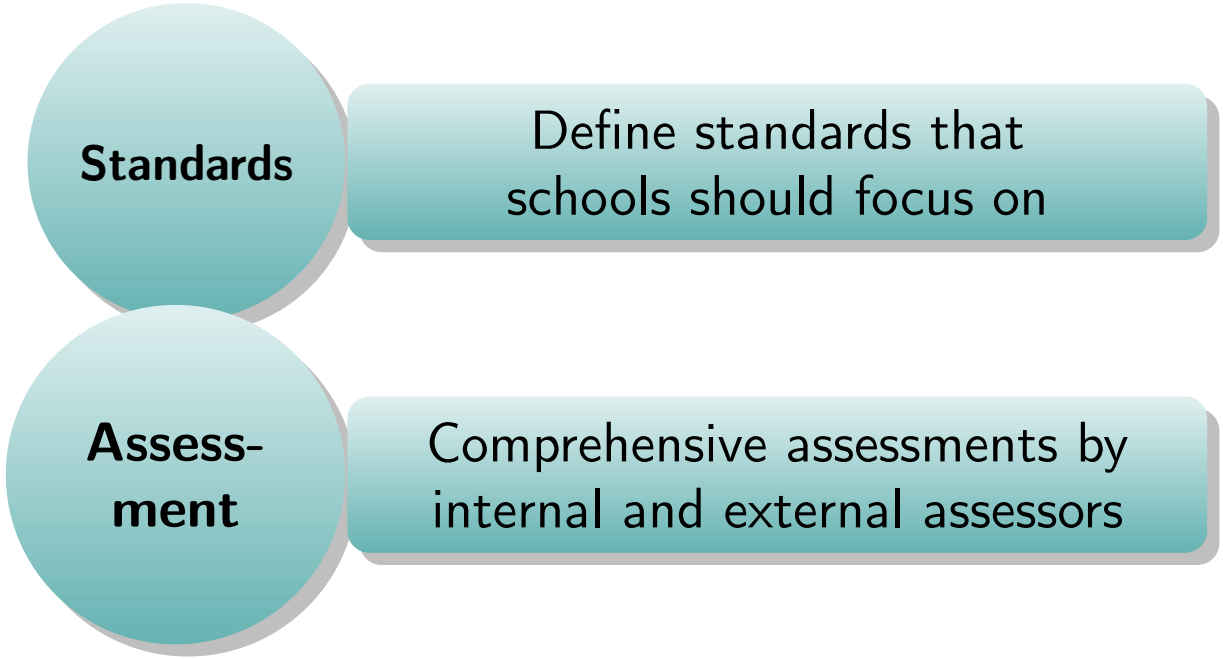
Program Components

Program Components

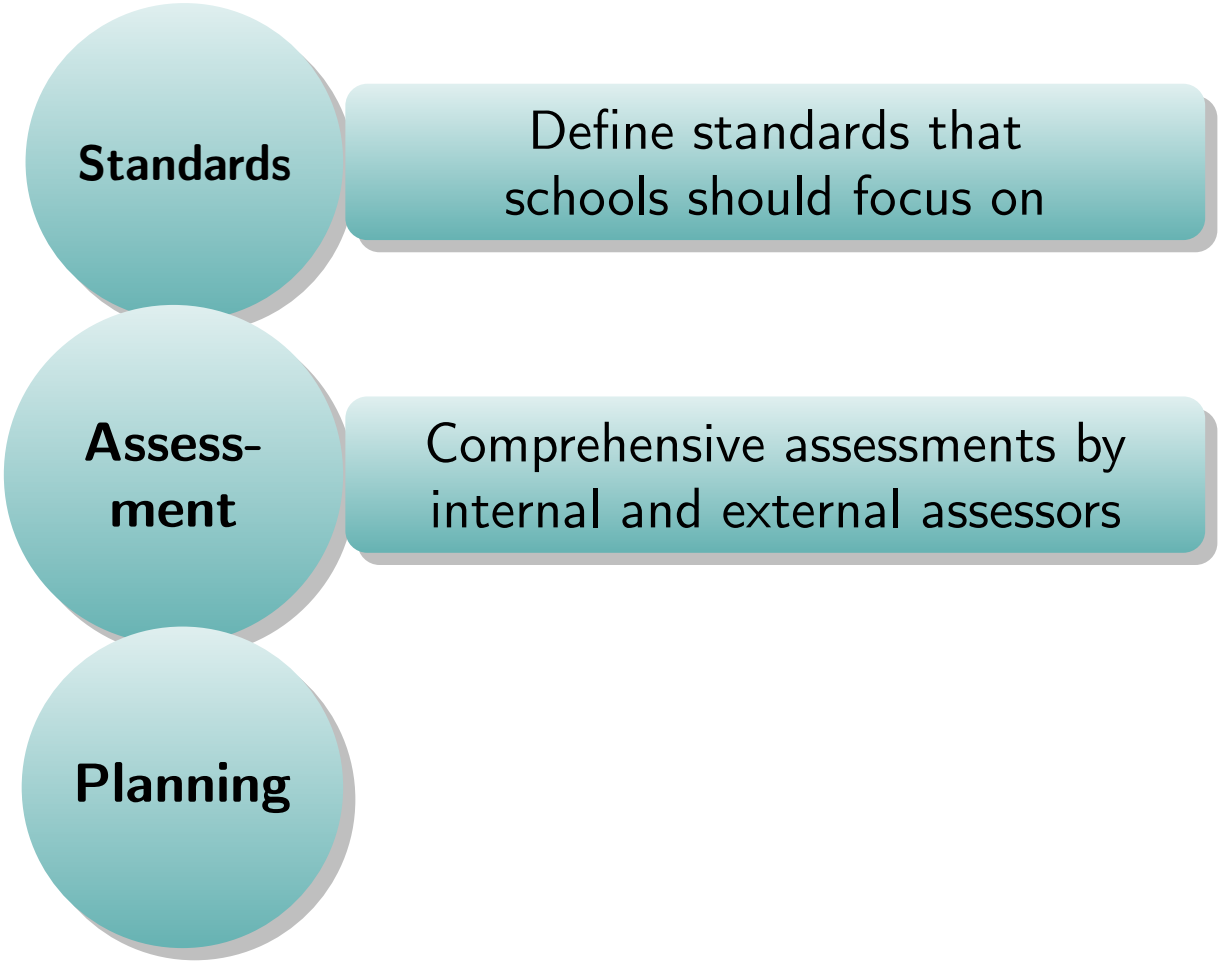
Program Components



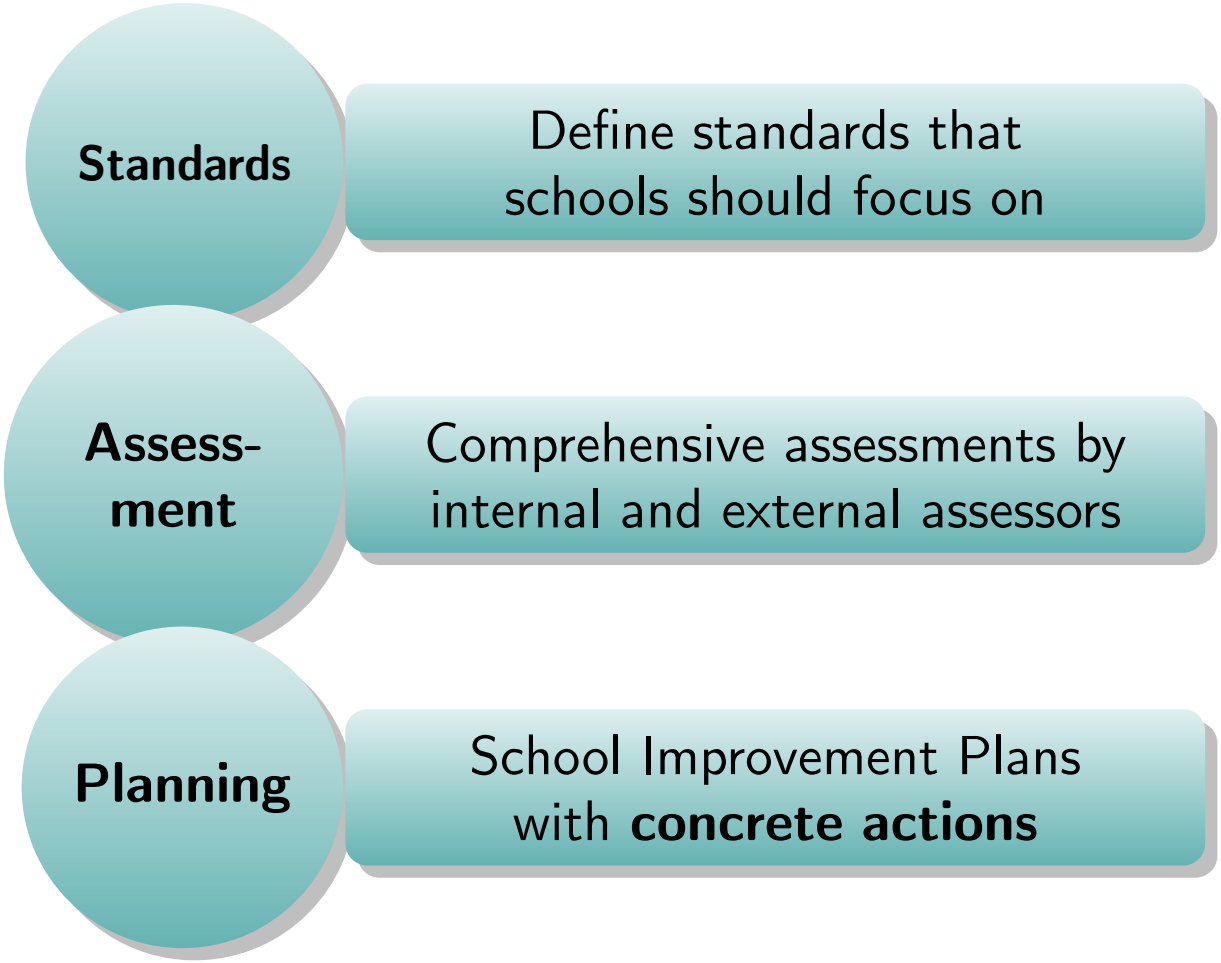
Program Components



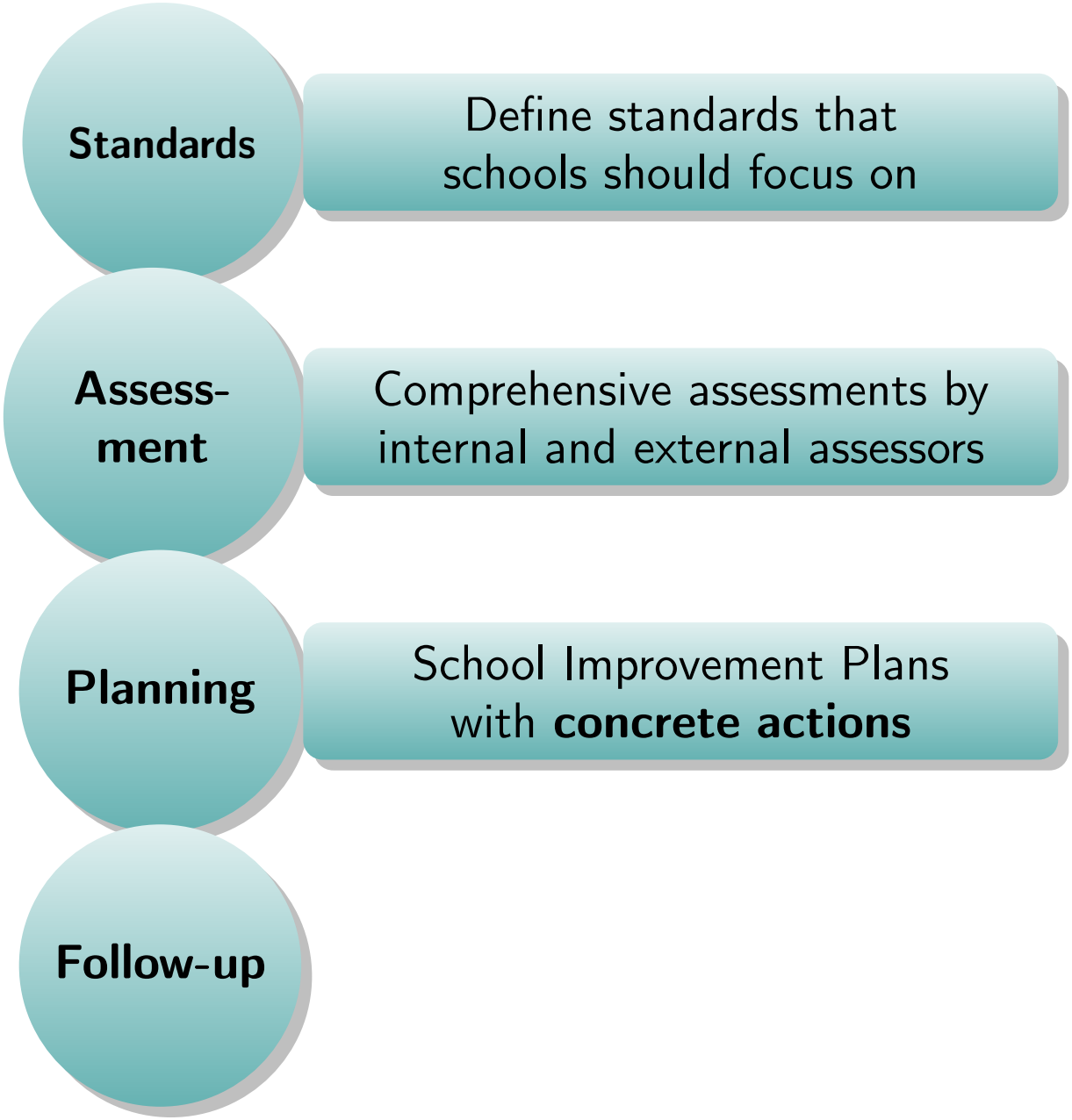
Program Components



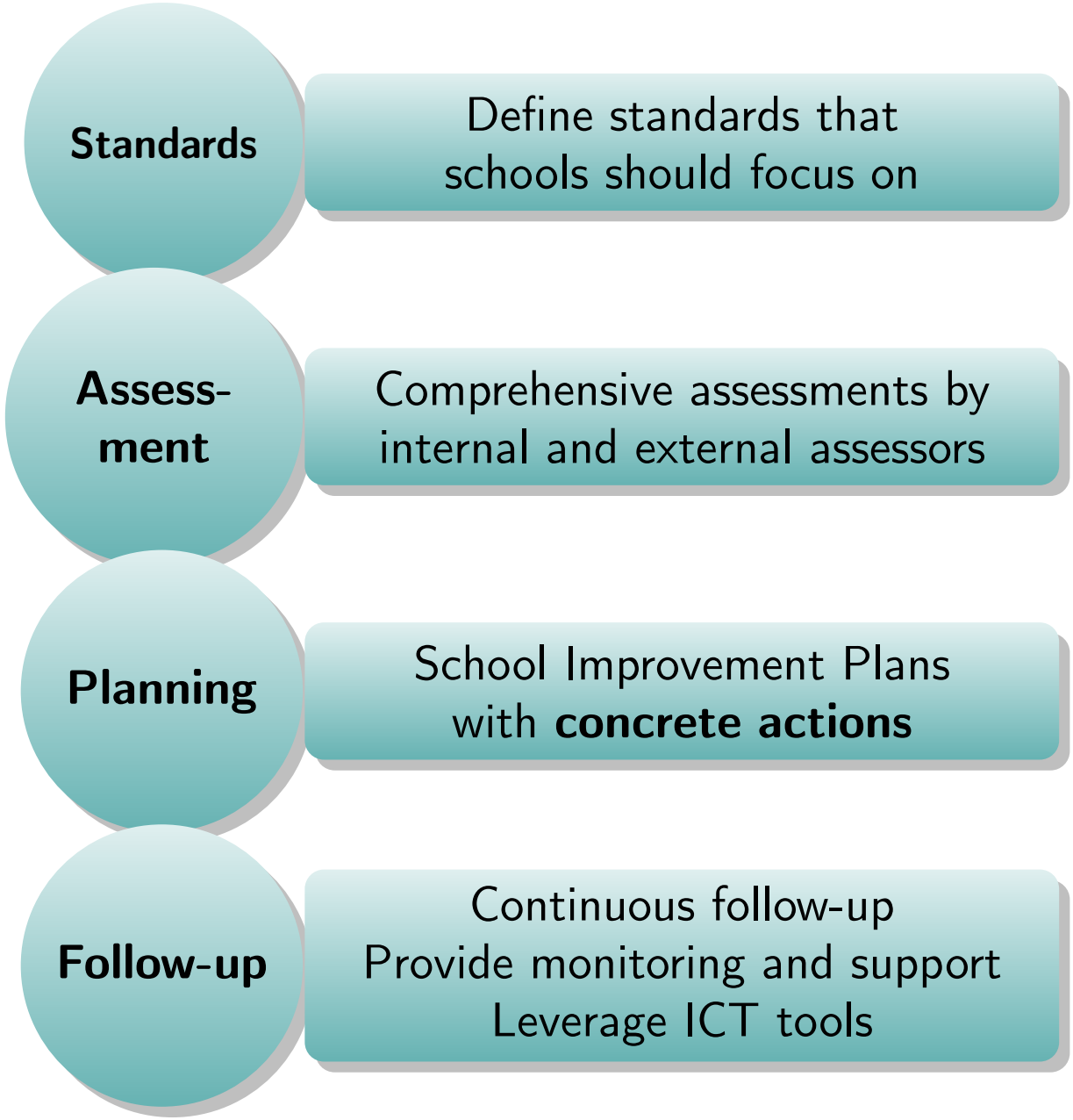
Program Components



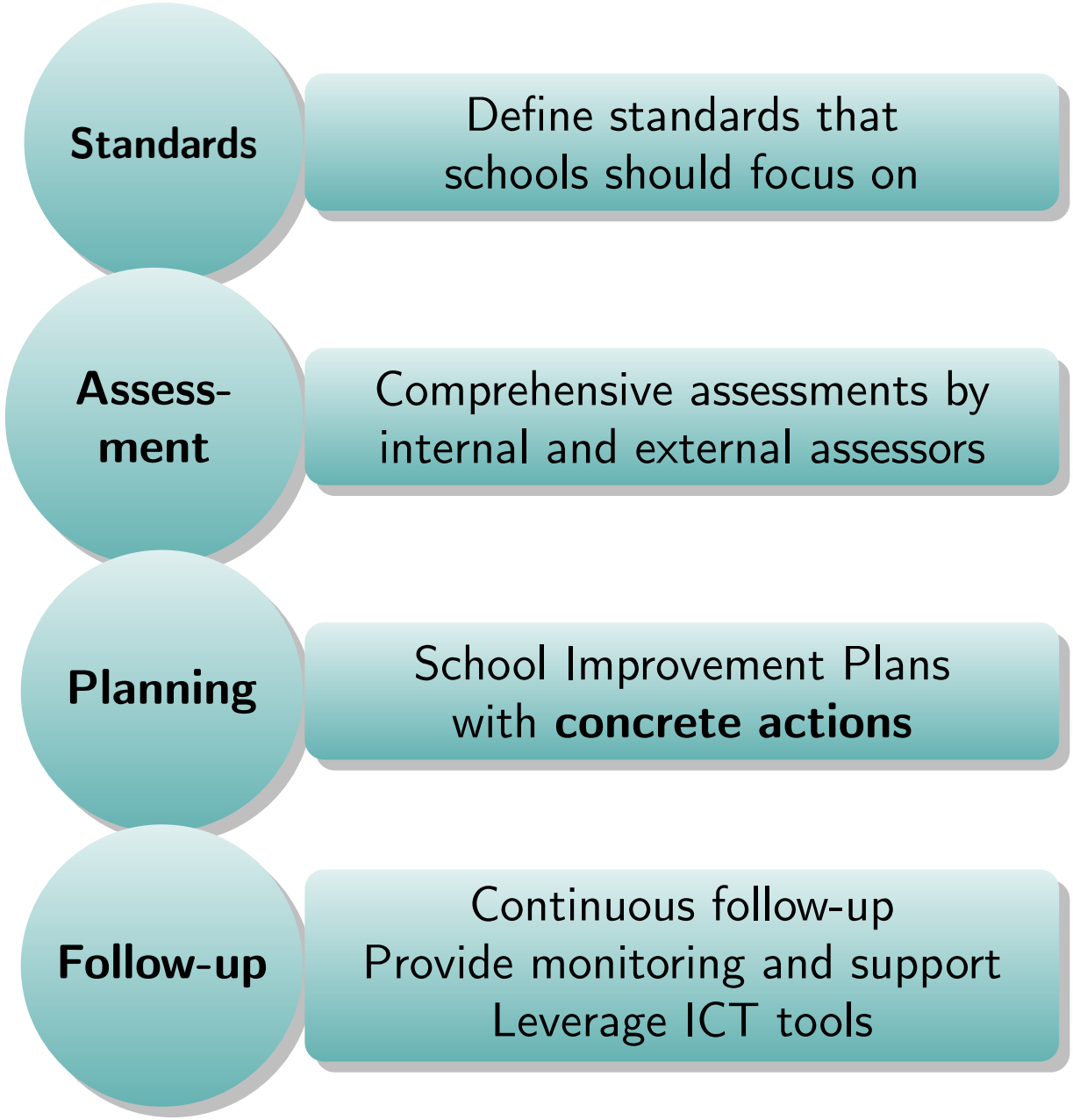
Program Components



Program Components

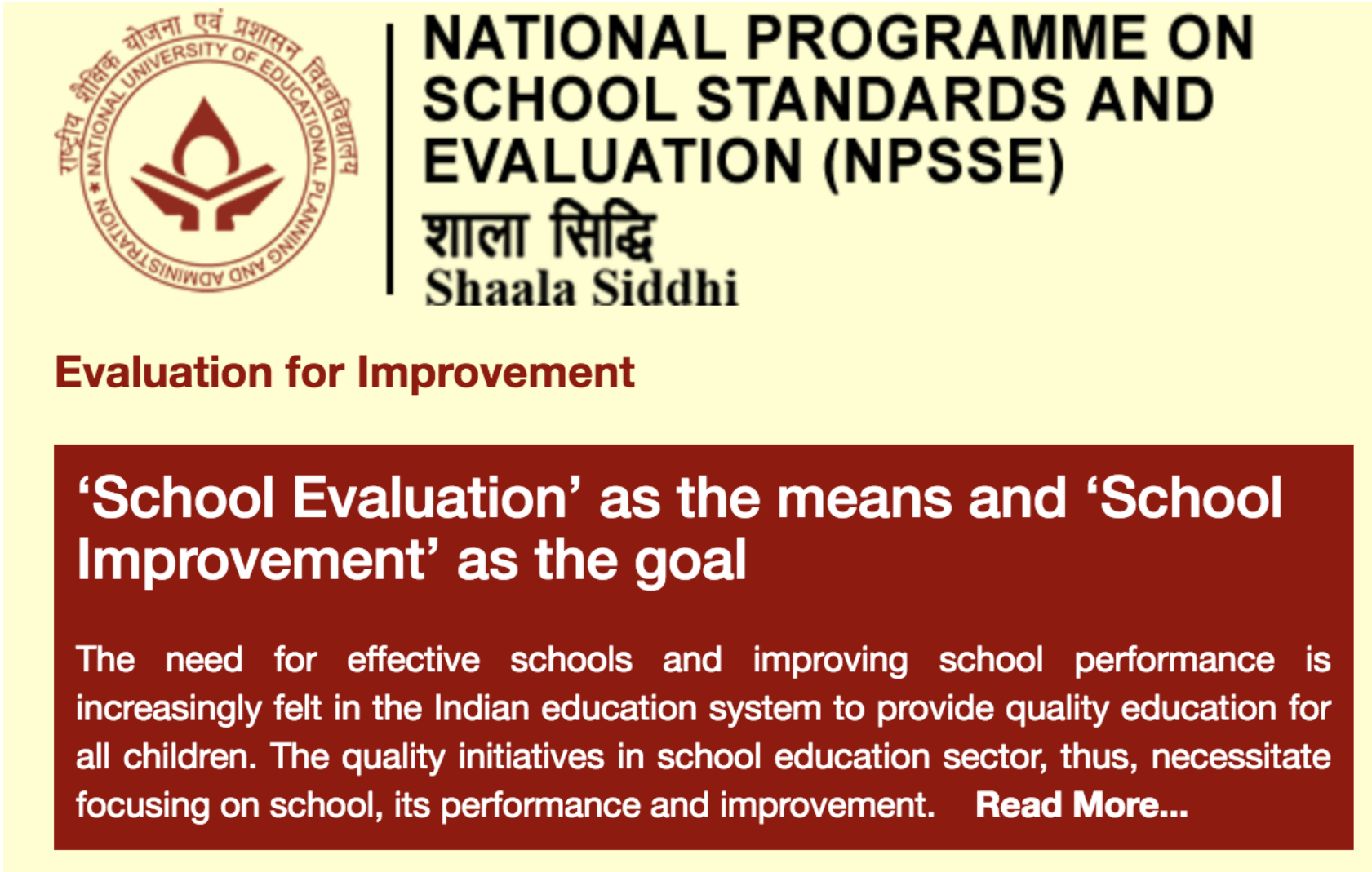


Program Components



This is very similar to a (since introduced) national policy

With a target of reaching 1.6 million schools!



Experimental design

- **Intervention scale:**
  - 2014-16: ~2000 schools in 5 districts, randomly assigned

Experimental design

- **Intervention scale:**
  - 2014-16: ~2000 schools in 5 districts, randomly assigned
- **Randomization:** Program assigned at **academic cluster level** (JSK, ~40 schools on average)
  - Program assignment at JSK level to avoid spillovers
  - Implemented in 153 JSKs; 51 to treatment, rest to control
  - All schools in the treatment JSK covered by the intervention
    - 1890 schools in treatment (1774 elementary, 116 secondary)
    - 3994 schools in control (3661 elementary, 273 control)

Experimental design

- **Intervention scale:**
  - 2014-16: ~2000 schools in 5 districts, randomly assigned
- **Randomization:** Program assigned at **academic cluster level** (JSK, ~40 schools on average)
  - Program assignment at JSK level to avoid spillovers
  - Implemented in 153 JSKs; 51 to treatment, rest to control
  - All schools in the treatment JSK covered by the intervention
    - 1890 schools in treatment (1774 elementary, 116 secondary)
    - 3994 schools in control (3661 elementary, 273 control)
- **Evaluation at scale:** intervention schools are a representative subset of the population of ~12k schools in these districts

Results

Four main results:

1. **Near-universal implementation of the assessments**
   - ~93% of elementary schools, 89% of secondary schools in treatment group were covered by the assessments and had SIPs prepared

Results

Four main results:

1. **Near-universal implementation of the assessments**
   - ~93% of elementary schools, 89% of secondary schools in treatment group were covered by the assessments and had SIPs prepared
2. **The assessments were meaningful**
   - Contain meaningful variation across schools
   - Most schools rated inadequate
   - Predicts *future* value-added and *future* teacher absence

Results

Four main results:

1. **Near-universal implementation of the assessments**
   - ~93% of elementary schools, 89% of secondary schools in treatment group were covered by the assessments and had SIPs prepared
2. **The assessments were meaningful**
   - Contain meaningful variation across schools
   - Most schools rated inadequate
   - Predicts *future* value-added and *future* teacher absence
3. **Yet, there was no change in the functioning of schools**
   - No increased monitoring and accountability
   - No change in classroom practice

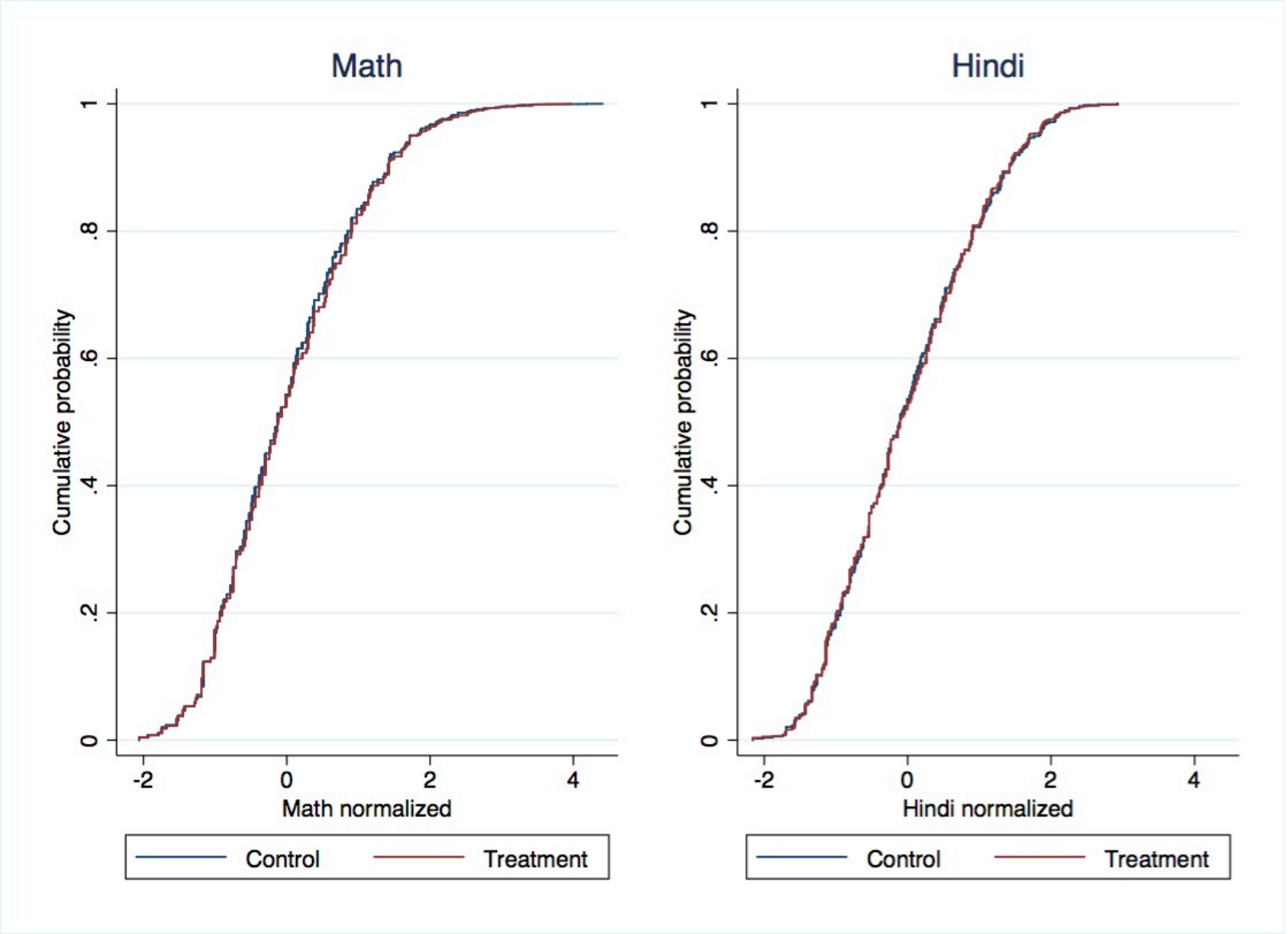
Results

Four main results:

1. **Near-universal implementation of the assessments**
   - ~93% of elementary schools, 89% of secondary schools in treatment group were covered by the assessments and had SIPs prepared
2. **The assessments were meaningful**
   - Contain meaningful variation across schools
   - Most schools rated inadequate
   - Predicts *future* value-added and *future* teacher absence
3. **Yet, there was no change in the functioning of schools**
   - No increased monitoring and accountability
   - No change in classroom practice
4. **Also, no change in test scores**
   - At 2, 14, 18 months!
   - Across administrative tests or our own

No effect on test scores

Student-level test scores



Note: Results on individual test scores in sample elementary schools from independent test data collection

Policy Implications

- Our results are relevant for informing the design of school governance programs in LMICs more generally

Policy Implications

- Our results are relevant for informing the design of school governance programs in LMICs more generally
- Similar program now mandated nationally, also of interest in other developing countries
  - Just this "pilot" exercise represents min. 27,000 man-days of work (13.5 man-days per school)
  - Only including days for assessment, SIP preparation by assessors and school staff alone
  - Universalizing just in MP implies **55x multiplication** in time spent (**~1.5 mn man-days**)
  - Excludes all logistical costs, opportunity costs for senior management, crowding out of other reforms

Reform 3: Personalized computer-aided instruction

Using technology to "Teach at the Right Level"

- One option that excites policy-makers is education technology

Reform 3: Personalized computer-aided instruction

Using technology to "Teach at the Right Level"

- One option that excites policy-makers is education technology
- In 2015, we evaluated a blended learning program (*Mindspark*)
  - Developed by a leading Indian education firm over a decade
    - Over 45,000 question Item Bank, used by over 400,000 students, administering over a million questions daily
  - **Individual, dynamically updated, assessment and content**
  - Instruction is targeted at children's actual level of achievement, **not the curriculum-mandated level**

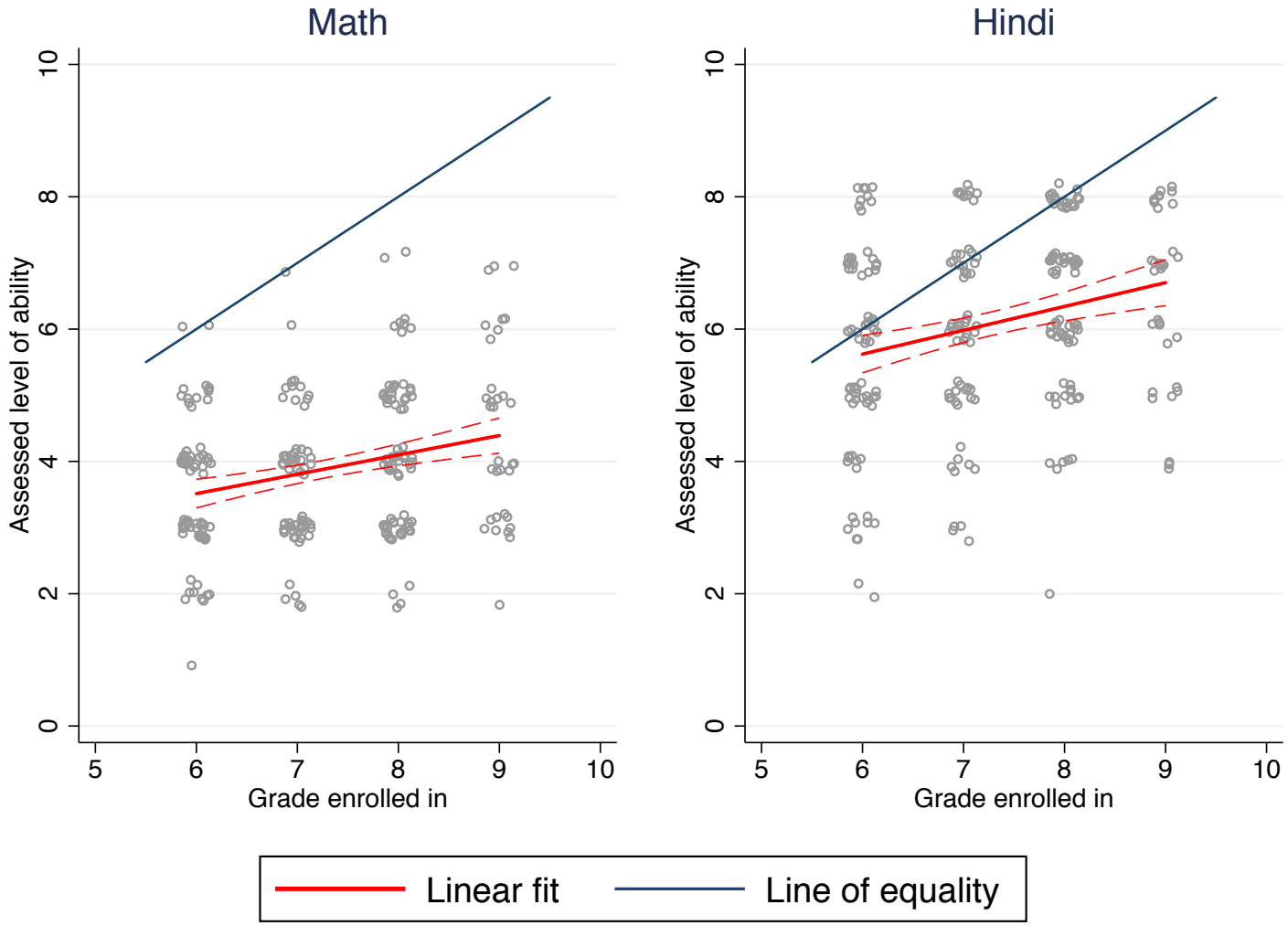
Reform 3: Personalized computer-aided instruction

Using technology to "Teach at the Right Level"

- One option that excites policy-makers is education technology
- In 2015, we evaluated a blended learning program (*Mindspark*)
  - Developed by a leading Indian education firm over a decade
    - Over 45,000 question Item Bank, used by over 400,000 students, administering over a million questions daily
  - **Individual, dynamically updated, assessment and content**
  - Instruction is targeted at children's actual level of achievement, **not the curriculum-mandated level**
- 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
  - all students from government secondary schools in Delhi

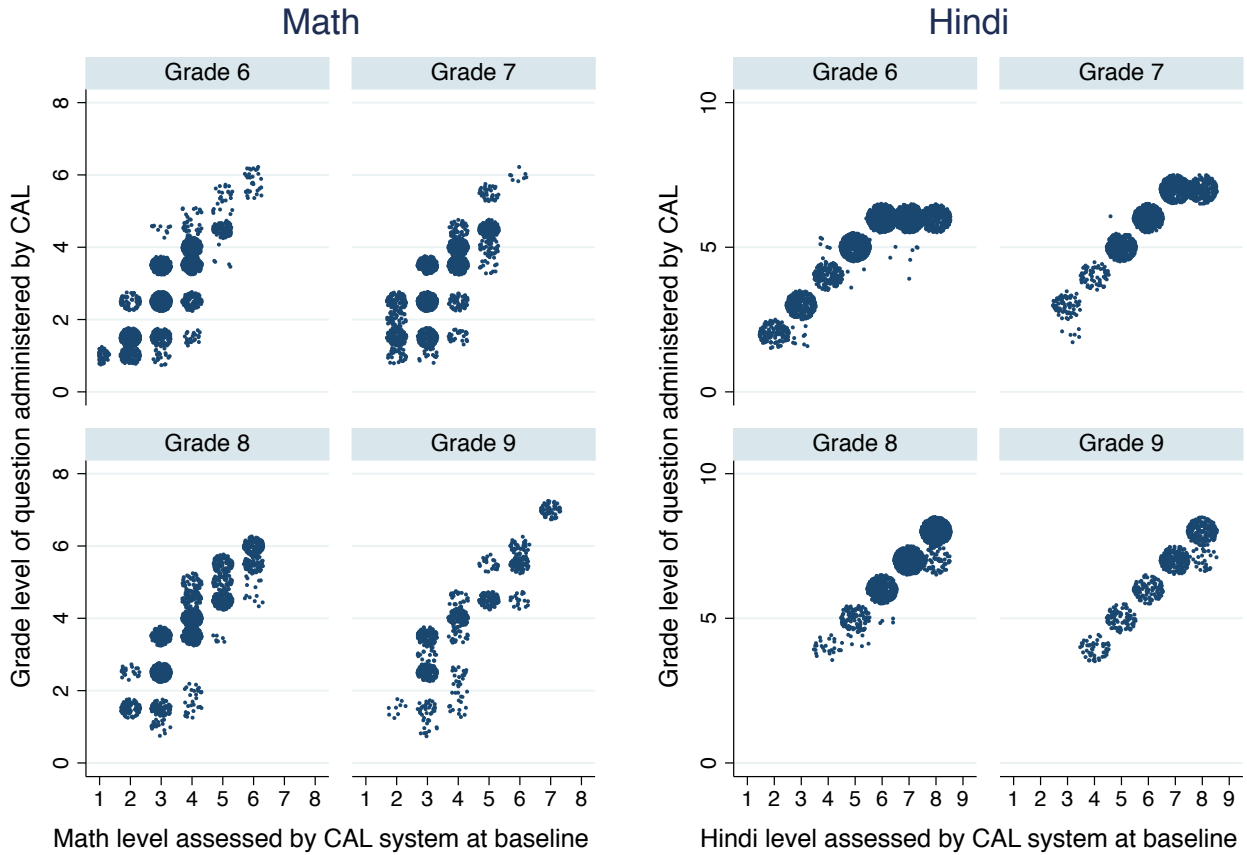
The problem

Low and dispersed achievement, mismatch with curriculum



This figure shows, for treatment group, the estimated initial level of student achievement (determined by the Mindspark CAL program) plotted against the grade they are enrolled in.

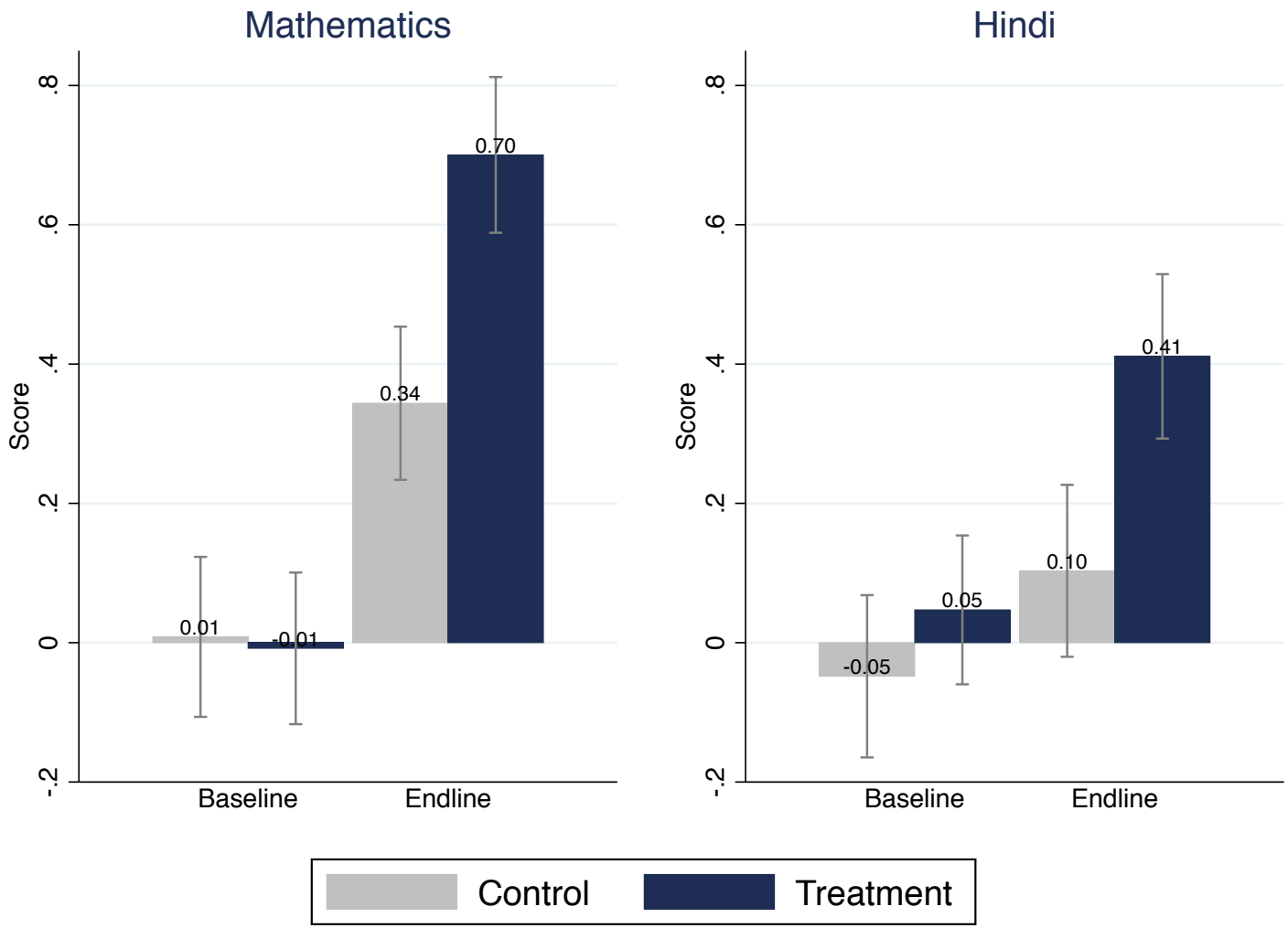
The potential of CAL to personalize instruction



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.

Results from out-of-school model

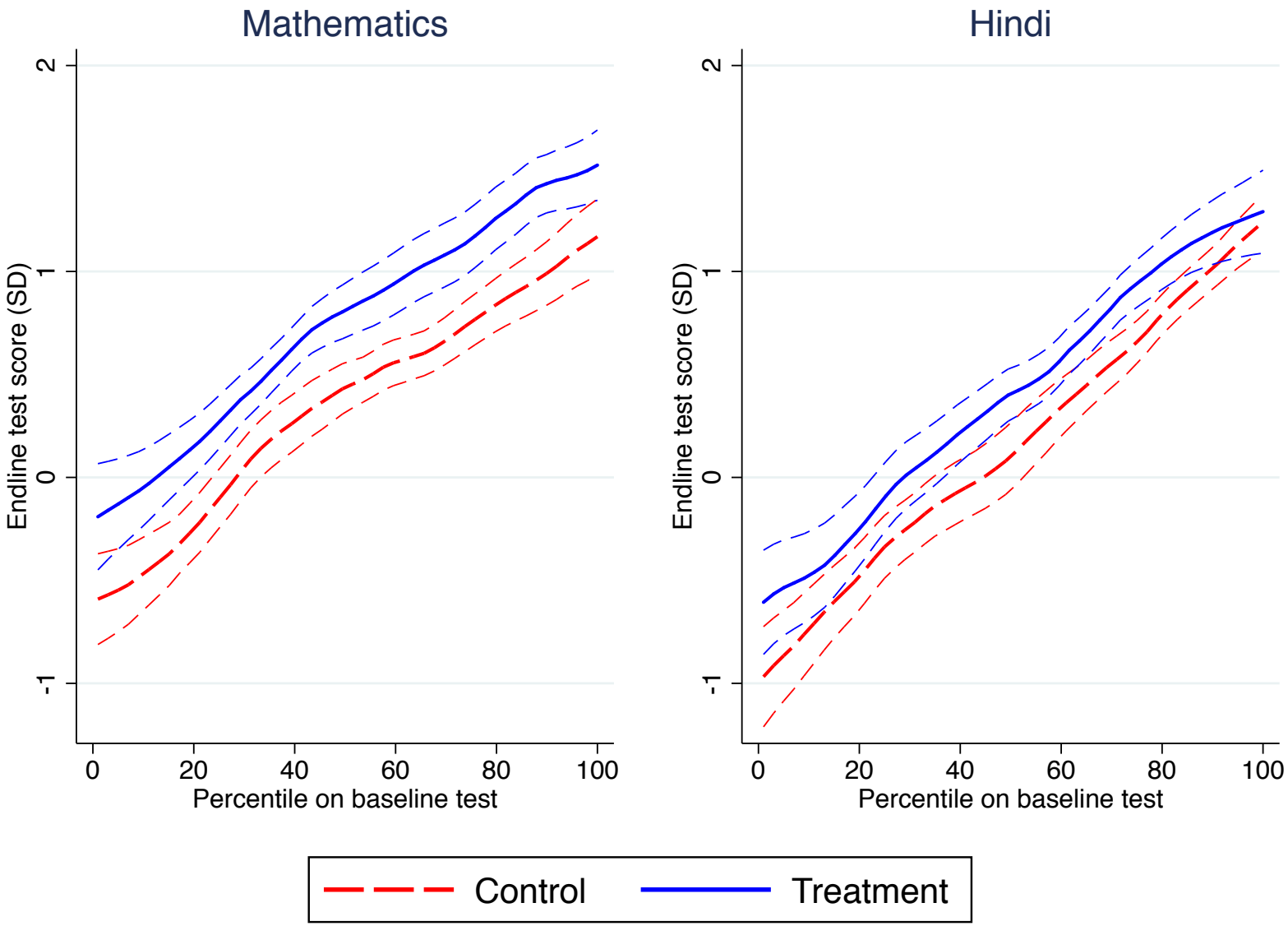
Very large impacts in Delhi



Muralidharan, Singh and Ganimian (2016)

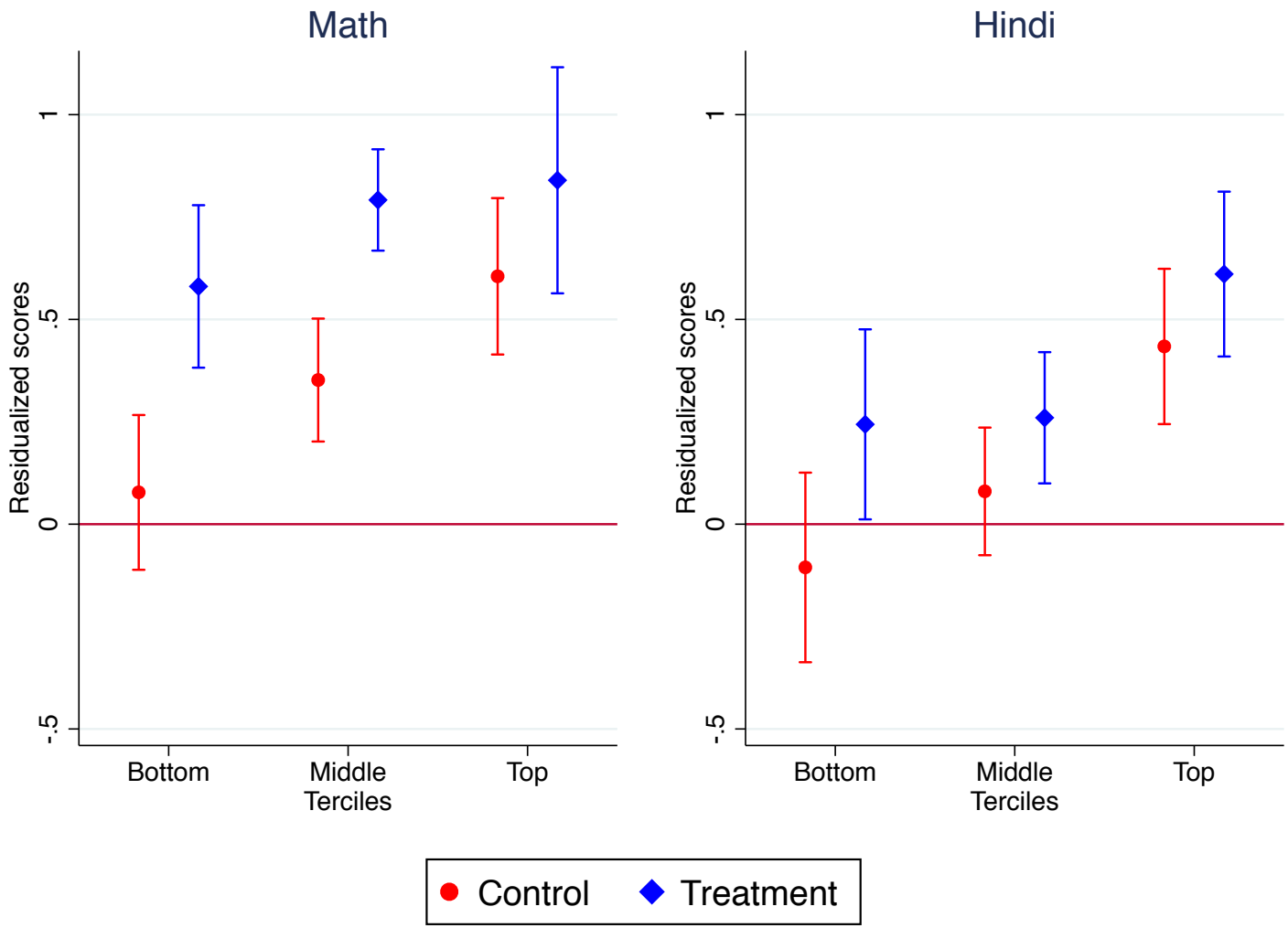
Results from out-of-school model

Unlike business-as-usual, the intervention taught *all* students



Results from out-of-school model

Unlike business-as-usual, the intervention taught *all* students



Muralidharan, Singh and Ganimian (2016)

Results Speak to Several Policy Debates

- **Technology in Education:**

Results Speak to Several Policy Debates

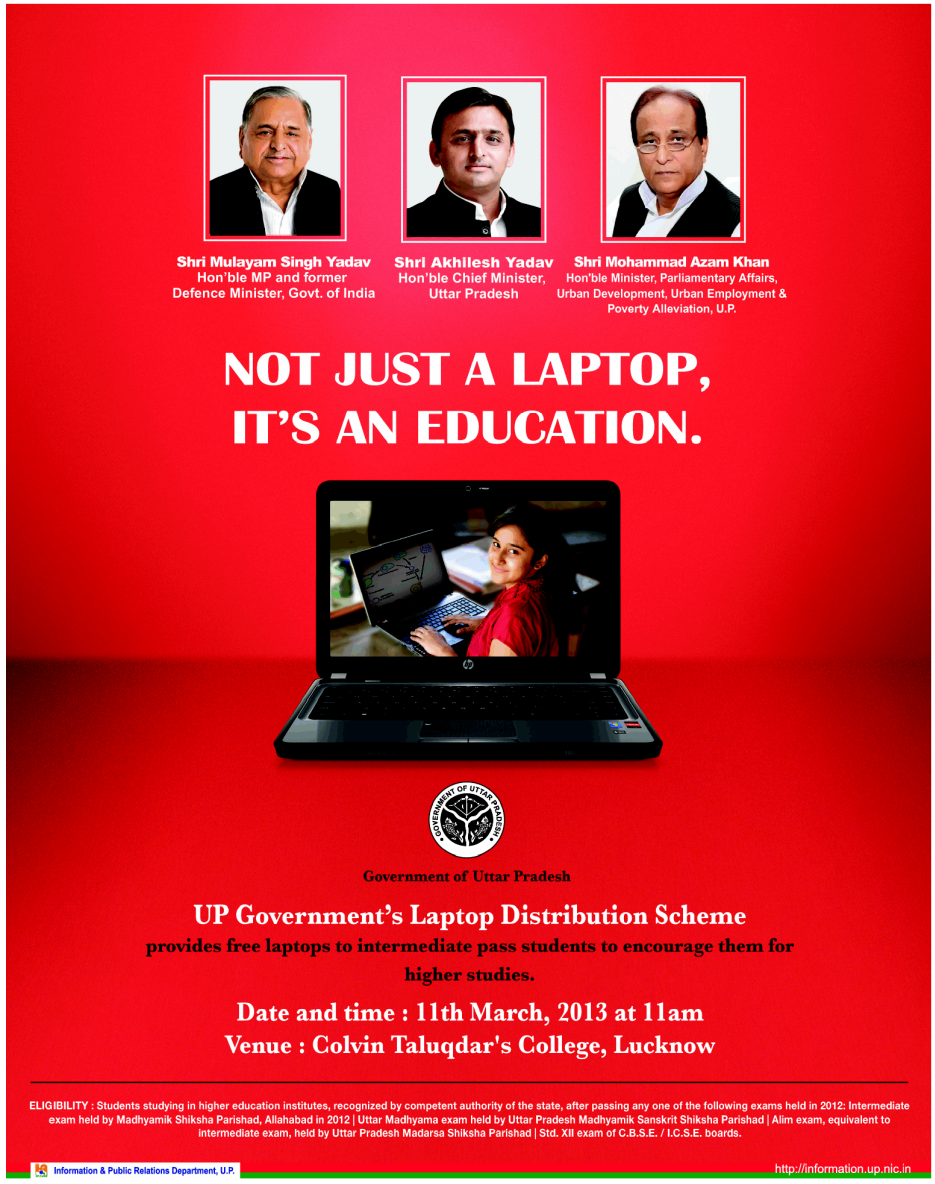
- **Technology in Education:**
  - Providing hardware alone seems to have very little impact on learning outcomes (Angrist and Lavy 2004; Malamud and Pop-Eleches 2011; Cristia et al. 2012; Beuermann et al. 2015)
  - Modest positive effects of CAL programs that focus on grade-appropriate instruction/review (Carillo et al. 2011; Lai et al. 2012, 2013, 2015; Mo et al. 2014a, b)
  - Large returns to customization (Banerjee et al. 2007)

Results Speak to Several Policy Debates

- **Technology in Education:**
  - Providing hardware alone seems to have very little impact on learning outcomes (Angrist and Lavy 2004; Malamud and Pop-Eleches 2011; Cristia et al. 2012; Beuermann et al. 2015)
  - Modest positive effects of CAL programs that focus on grade-appropriate instruction/review (Carillo et al. 2011; Lai et al. 2012, 2013, 2015; Mo et al. 2014a, b)
  - Large returns to customization (Banerjee et al. 2007)
- **Developing country education**
  - Mismatch between student preparedness and curriculum is a major concern (Banerjee and Duflo, 2012; Pritchett and Beatty 2015)
  - May explain low productivity of education in these settings
  - "Teaching at the right level" effective in raising achievement in primary schools (Banerjee et al 2007)
  - But unclear whether feasible at secondary school level

Understanding political demand for computers in education

Understanding political demand for computers in education

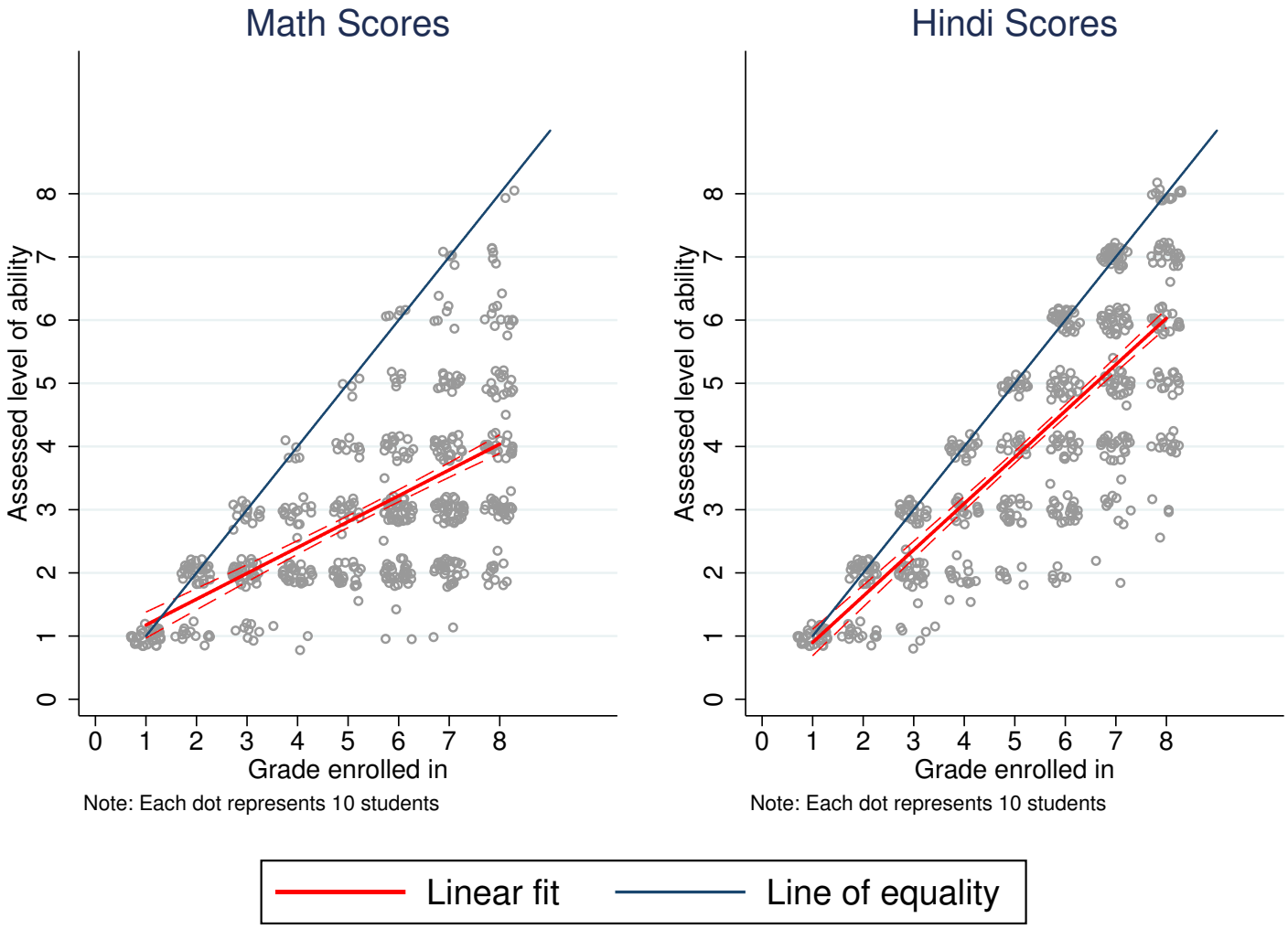


CAL at policy relevant scales

Evaluating Mindspark in government schools

- Our Delhi study is best regarded as an "efficacy trial"
  - delivered at small scale, with high fidelity, by motivated staff
- The pathway to scale will lie in extending to govt schools
  - substantial policy demand for tech in education
    - typically with little idea of how tech will lead to more learning
    - if successful, might be rapidly scaleable
- Delivering in govt schools non-trivially different
  - working with school teachers to deliver intervention
  - integrating CAL in timetable
    - special adaptation due to hardware constraints
- We are working with GoR and EI for introduction in the coming school year.
  - ~6000 students in treatment across 40 schools
  - urban and rural areas

An initial assessment of heterogeneity



Muralidharan and Singh (ongoing)

Summary

An evidence-based approach to education reform

- Policy efforts at increasing **learning in schools** have been much less successful than increasing participation
- Governments are now increasingly aware of this issue and willing to consider measures

Summary

An evidence-based approach to education reform

- Policy efforts at increasing **learning in schools** have been much less successful than increasing participation
- Governments are now increasingly aware of this issue and willing to consider measures
- We need to know whether these work in terms of outcomes
  - just having delivered a program is not enough
  - important for understanding which approaches should be continued, which should be modified, which are best left alone
  - too many fads in education, need discipline from evaluations
  - often requires flexibility in methods (descriptive, experimental + non-experimental)
  - there are things to be learned even from null results

Summary

An evidence-based approach to education reform

- Policy efforts at increasing **learning in schools** have been much less successful than increasing participation
- Governments are now increasingly aware of this issue and willing to consider measures
- We need to know whether these work in terms of outcomes
  - just having delivered a program is not enough
  - important for understanding which approaches should be continued, which should be modified, which are best left alone
  - too many fads in education, need discipline from evaluations
  - often requires flexibility in methods (descriptive, experimental + non-experimental)
  - there are things to be learned even from null results
- With the right partnerships between policymakers and academics, this is possible.