

Adapting for scale:

Experimental Evidence on computer-aided instruction in India

Karthik Muralidharan

UCSD, NBER, JPAL

Abhijeet Singh

SSE, J-PAL, CEPR

WWHGE Conference, Oxford

September 2026

Introduction

- ▶ The rapid growth of RCTs in economics has often been motivated by trying to find “what works” and then scaling it up
- ▶ Yet, globally, many interventions shown to be effective in small-scale efficacy trials are being found to be much less effective at scale ([Linos et al., 2022](#); [DellaVigna & Linos, 2022](#); [Kraft et al., 2024](#); [Mitchell et al., 2023](#); [Araujo et al., 2016](#))
 - ▶ Plausibly reflects constraints that bind at scale but did not at “proof-of-concept” (efficacy) stage

Introduction

- ▶ The rapid growth of RCTs in economics has often been motivated by trying to find “what works” and then scaling it up
- ▶ Yet, globally, many interventions shown to be effective in small-scale efficacy trials are being found to be much less effective at scale ([Linos et al., 2022](#); [DellaVigna & Linos, 2022](#); [Kraft et al., 2024](#); [Mitchell et al., 2023](#); [Araujo et al., 2016](#))
 - ▶ Plausibly reflects constraints that bind at scale but did not at “proof-of-concept” (efficacy) stage
- ▶ Thus, delivering on the promising of “evidence-based” policy requires as much (or more) attention to adapting successful programs for scalability and testing (intermediate) scalable models
 - ▶ Especially important given NGO and researcher incentives to find positive results

“... short-term incentives encourage researchers to create a Petri dish that provides results that are overly optimistic: consumers of scientific journals demand studies that report important insights [...] The result is that we are essentially performing efficacy tests on steroids without telling outsiders.” ([List, 2024](#))

Two ways to study scaling (and a missing middle)

► **Scaling as replication with larger N and different settings**

- Dominant approach to scaling after establishing efficacy (e.g. in clinical trials)
- Scaling problem is seen as one of maintaining compliance to validated protocols at large scale
- **Problem:** Scale-ups are rarely evaluated; and those that are are often ineffective at scale (List, 2022)
- **Challenge:** Most social programs require adaptation to address new constraints that bind at scale
- This challenge is **distinct** from external validity across contexts (as addressed by, e.g., [Banerjee, Duflo, et al. \(2015\)](#), [Banerjee, Karlan, & Zinman \(2015\)](#))

Two ways to study scaling (and a missing middle)

► **Scaling as replication with larger N and different settings**

- Dominant approach to scaling after establishing efficacy (e.g. in clinical trials)
- Scaling problem is seen as one of maintaining compliance to validated protocols at large scale
- **Problem:** Scale-ups are rarely evaluated; and those that are are often ineffective at scale (List, 2022)
- **Challenge:** Most social programs require adaptation to address new constraints that bind at scale
- This challenge is **distinct** from external validity across contexts (as addressed by, e.g., [Banerjee, Duflo, et al. \(2015\)](#), [Banerjee, Karlan, & Zinman \(2015\)](#))

► **Experimentation at scale** ([Muralidharan & Niehaus \(2017\)](#))

- Induce randomized variation for programs at scale, in representative population and situation
- Results are directly policy-relevant, informative of welfare at scale ([Muralidharan et al. \(2016\)](#))
- But if programs don't work, limited insight on what **might** have made it work ([Muralidharan & Singh \(2020\)](#))

Two ways to study scaling (and a missing middle)

► **Scaling as replication with larger N and different settings**

- Dominant approach to scaling after establishing efficacy (e.g. in clinical trials)
- Scaling problem is seen as one of maintaining compliance to validated protocols at large scale
- **Problem:** Scale-ups are rarely evaluated; and those that are are often ineffective at scale (List, 2022)
- **Challenge:** Most social programs require adaptation to address new constraints that bind at scale
- This challenge is **distinct** from external validity across contexts (as addressed by, e.g., [Banerjee, Duflo, et al. \(2015\)](#), [Banerjee, Karlan, & Zinman \(2015\)](#))

► **Experimentation at scale** ([Muralidharan & Niehaus \(2017\)](#))

- Induce randomized variation for programs at scale, in representative population and situation
- Results are directly policy-relevant, informative of welfare at scale ([Muralidharan et al. \(2016\)](#))
- But if programs don't work, limited insight on what **might** have made it work ([Muralidharan & Singh \(2020\)](#))

► **This paper: Adapting interventions for scalability, *and* evaluating before scaling**

- The efficacy trial is “proof-of-concept” for the basic theory-of-change
- **Intermediate step:** Develop and test *adapted* variants that embed constraints that will bind at scale but preserve the (invariant) core of the program
- This can then provide a robust template for scaling if successful, or pause premature scaling if not (see, e.g., [List \(2024\)](#))

Context and Motivation

Provide this key intermediate step (of adaptation and evaluation) in the context of a highly promising intervention where we found large positive effects in an efficacy trial

- ▶ **Intervention:** Personalized computer-aided instruction (*Mindspark*) which targets teaching at the level of academic preparation
 - ▶ Address issues of “curriculum mismatch” and within-class heterogeneity

Context and Motivation

Provide this key intermediate step (of adaptation and evaluation) in the context of a highly promising intervention where we found large positive effects in an efficacy trial

- ▶ **Intervention:** Personalized computer-aided instruction (*Mindspark*) which targets teaching at the level of academic preparation
 - ▶ Address issues of “curriculum mismatch” and within-class heterogeneity
- ▶ **Efficacy trial:** We demonstrated large gains over 4-5 months of 0.22σ in language and 0.36σ in Math ([Muralidharan, Singh, & Ganimian, 2019](#))
 - ▶ 619 students, 314 treated (individually randomized)
 - ▶ Instruction after school, *supplements* instruction; small scale of three centres in Delhi
 - ▶ Motivated providers, high fidelity implementation; motivated self-selected study participants
 - ▶ Limited duration (half a school year)
 - ▶ Expensive in absolute terms (although cost-effective vs regular instruction)
 - ▶ **These results were a proof-of-concept, but not a template for scaling.**

Context and Motivation

Provide this key intermediate step (of adaptation and evaluation) in the context of a highly promising intervention where we found large positive effects in an efficacy trial

- ▶ **Intervention:** Personalized computer-aided instruction (*Mindspark*) which targets teaching at the level of academic preparation
 - ▶ Address issues of “curriculum mismatch” and within-class heterogeneity
- ▶ **Efficacy trial:** We demonstrated large gains over 4-5 months of 0.22σ in language and 0.36σ in Math ([Muralidharan, Singh, & Ganimian, 2019](#))
 - ▶ 619 students, 314 treated (individually randomized)
 - ▶ Instruction after school, *supplements* instruction; small scale of three centres in Delhi
 - ▶ Motivated providers, high fidelity implementation; motivated self-selected study participants
 - ▶ Limited duration (half a school year)
 - ▶ Expensive in absolute terms (although cost-effective vs regular instruction)
 - ▶ **These results were a proof-of-concept, but not a template for scaling.**
- ▶ **Pathways to scale need models for implementation in public schools**
 - ▶ Especially important for reaching underprivileged students
 - ▶ Also reflect a substantial share of public funds for EdTech
 - ▶ But that needs substantial adaptation to reflect new constraints, no guarantee of success.

This paper

Adapting and testing for scale in public schools

- ▶ **We developed an implementation model for public schools**
 - ▶ Implemented entirely within the school day without adding school time
 - ▶ Schedules 1 computer lab period every day for students to work on *Mindspark*
 - ▶ Replaces scheduled class time in Math/Hindi, and for remediation
 - ▶ Regular teacher expected to accompany students to the lab
 - ▶ With resource constraints, students paired to a single PC if needed
 - ▶ Program provided a lab-in-charge responsible for hardware, logistics
 - ▶ Dedicated school-specific resource for the first 2 years, reduced to 1/3 in Y3
- ▶ **This is a substantially different intervention from the efficacy trial**

This paper

Adapting and testing for scale in public schools

- ▶ **We developed an implementation model for public schools**
 - ▶ Implemented entirely within the school day without adding school time
 - ▶ Schedules 1 computer lab period every day for students to work on *Mindspark*
 - ▶ Replaces scheduled class time in Math/Hindi, and for remediation
 - ▶ Regular teacher expected to accompany students to the lab
 - ▶ With resource constraints, students paired to a single PC if needed
 - ▶ Program provided a lab-in-charge responsible for hardware, logistics
 - ▶ Dedicated school-specific resource for the first 2 years, reduced to 1/3 in Y3
 - ▶ **This is a substantially different intervention from the efficacy trial**
- ▶ **We evaluated this model for three years in a clustered RCT**
 - ▶ 80 schools in rural and urban Rajasthan state
 - ▶ ~20x the number of treated students per year
 - ▶ Covers all primary and middle school grades (from Classes 1-8)
 - ▶ Primary outcomes: independent measures of student achievement
 - ▶ Additionally, detailed data on usage, measures of classroom practice, student/teacher interviews
 - ▶ We focus on 18-month results first (with dedicated LIC) and then show results for Y3

Intervention Design

► **Schedule:**

- Instruction during the school day, replacing regular instruction
- ~ 1 class period (35-40 mins) scheduled per school day
- Divided between two subjects: Math and Hindi

► **Implementation:**

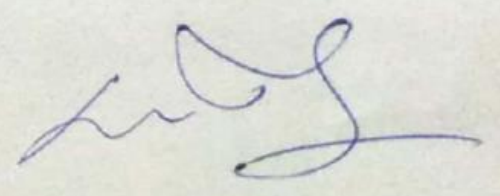
- Children go to a computer lab during the assigned Mindspark period
- Where feasible, each child works on their own computer (1:1)
- If more students than computers, students work in assigned pairs (2:1)
- Students in a pair are matched based on their diagnosed level of achievement

► **Personnel:**

- Implementation carried out with regular subject teachers (who were trained)
- In addition, a lab-in-charge to take care of the computer lab, generate Mindspark usernames, manage program logistics etc
- **Reduced to 2-days-per-week in Y3**

Time-tabling Mindspark

GSSS I : Mindspark Time Table						
PERIOD	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
1						
2	STD-6 (HINDI)	STD-6 (HINDI)	STD-6 (HINDI)	STD-5 (MATH)	STD-5 (MATH)	STD-5 (MATH)
3	STD-5 (HINDI)	STD-5 (HINDI)	STD-5 (HINDI)	STD-8 (HINDI)	STD-8 (HINDI)	STD-8 (HINDI)
4	STD-1&2 (MATH)	STD-1&2 (MATH)	STD-1&2 (MATH)	STD-4 (HINDI)	STD-4 (HINDI)	STD-4 (HINDI)
5	STD-3 (MATH)	STD-3 (MATH)	STD-3 (MATH)	STD-1&2 (HINDI)	STD-1&2 (HINDI)	STD-1&2 (HINDI)
6	STD-4 (MATH)	STD-4 (MATH)	STD-4 (MATH)	STD-6 (MATH)	STD-6 (MATH)	STD-6 (MATH)
7	STD-7 (MATH)	STD-7 (MATH)	STD-7 (MATH)	STD-7 (HINDI)	STD-7 (HINDI)	STD-7 (HINDI)
8	STD-8 (MATH)	STD-8 (MATH)	STD-8 (MATH)	STD-3 (HINDI)	STD-3 (HINDI)	STD-3 (HINDI)


प्रधानाचार्य

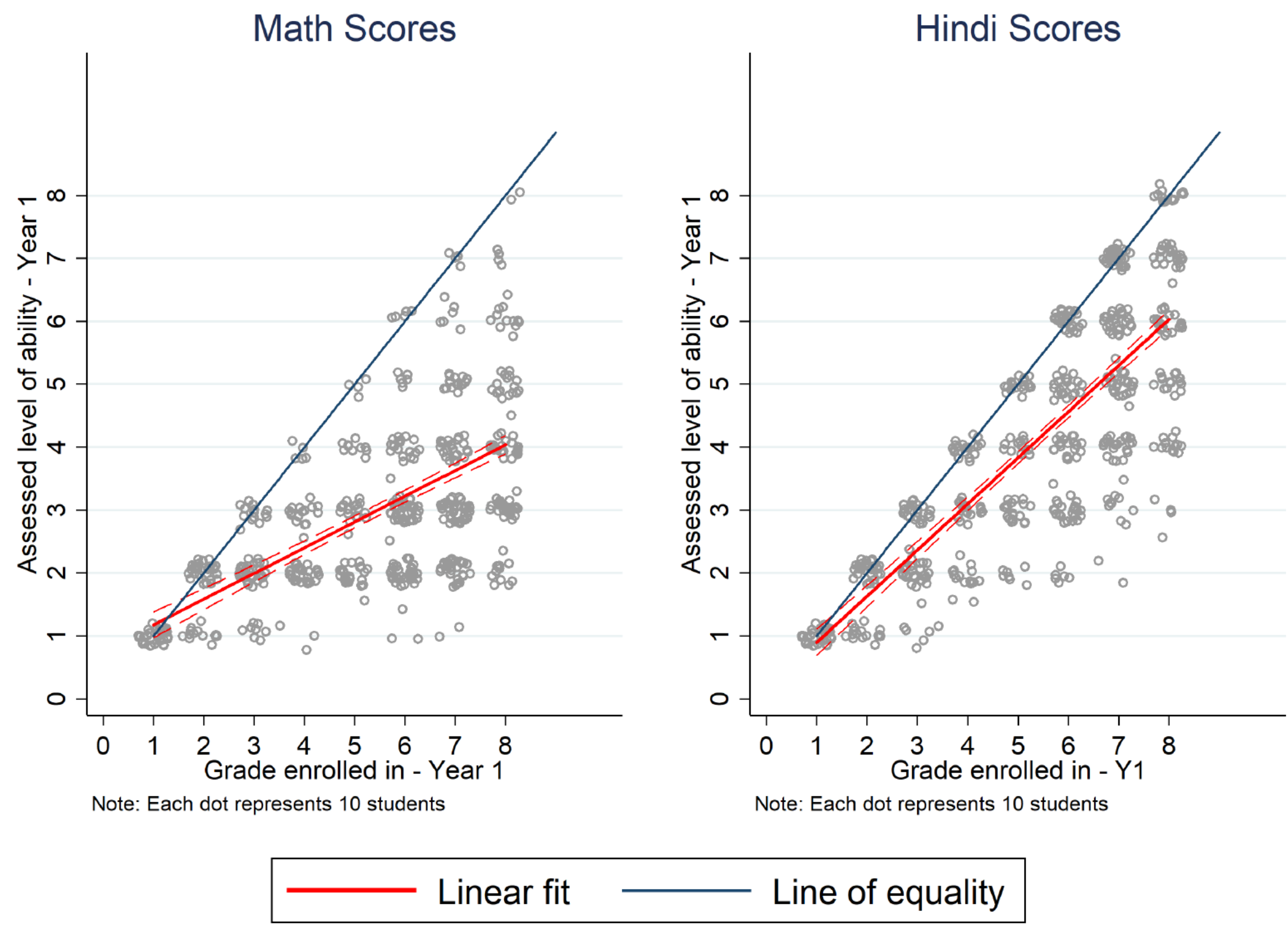
Experiment Design

- ▶ **Geographical scope:** Four districts of Rajasthan
 - ▶ Churu, Jhunjhunun, Udaipur, Dungarpur
 - ▶ across both urban and rural areas
- ▶ **Schools:** Adarsh schools which have Grades 1-10
 - ▶ 80 schools split between treatment and control
 - ▶ **pairwise randomization**, stratified within district
 - ▶ These are integrated (*Adarsh*) schools from Grade 1 to Grade 10 or 12
 - ▶ Our sample schools are fairly typical of such integrated public schools
- ▶ **Students:** All students in Grades 1-8 assigned to Mindspark
 - ▶ ~6500 students per year with access to treatment
 - ▶ ~20x the size of the Delhi trial

Geographical scope



Actual vs. expected levels



Effects in the full sample

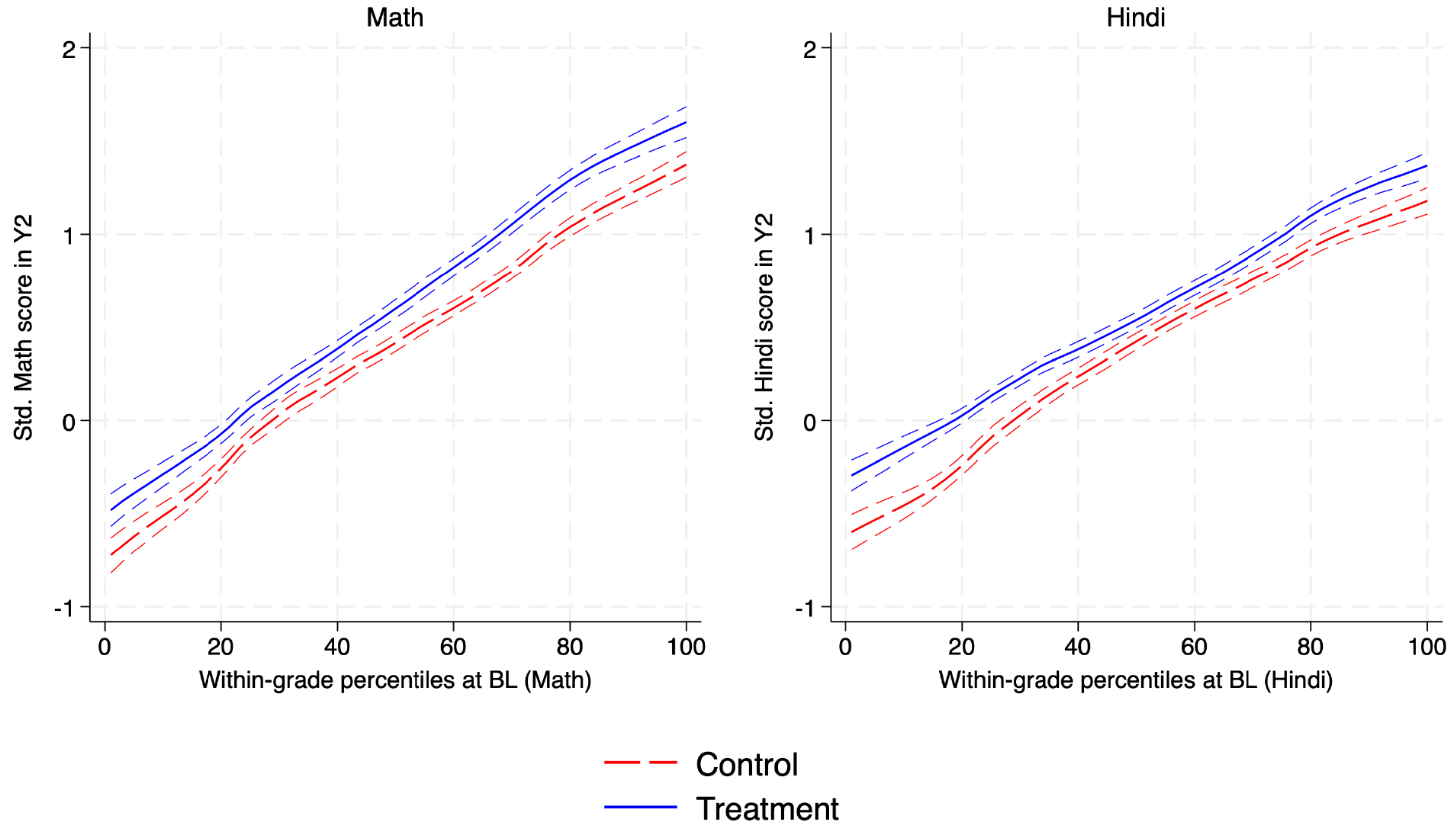
Year-specific, pooling all grades

	Math		Hindi	
	Year 1	Year 2	Year 1	Year 2
Treatment	0.15*** (0.032)	0.22*** (0.036)	0.11*** (0.026)	0.20*** (0.032)
Baseline score	0.67*** (0.010)	0.59*** (0.016)	0.72*** (0.013)	0.65*** (0.015)
Observations	7994	8733	7994	8733
R-squared	0.531	0.346	0.505	0.383

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Standard errors are clustered at the stratum level. This table presents Intention-to-treat treatment effects of the program at the end of ~6 months of treatment (EL Y1) and 18 months (EL Y2). Test scores are standardized to have a mean of zero and a standard deviation of one in the baseline in grade 5 in the control group. All regressions include strata (school-pair) fixed effects and student gender.

Gains across the distribution of baseline test scores

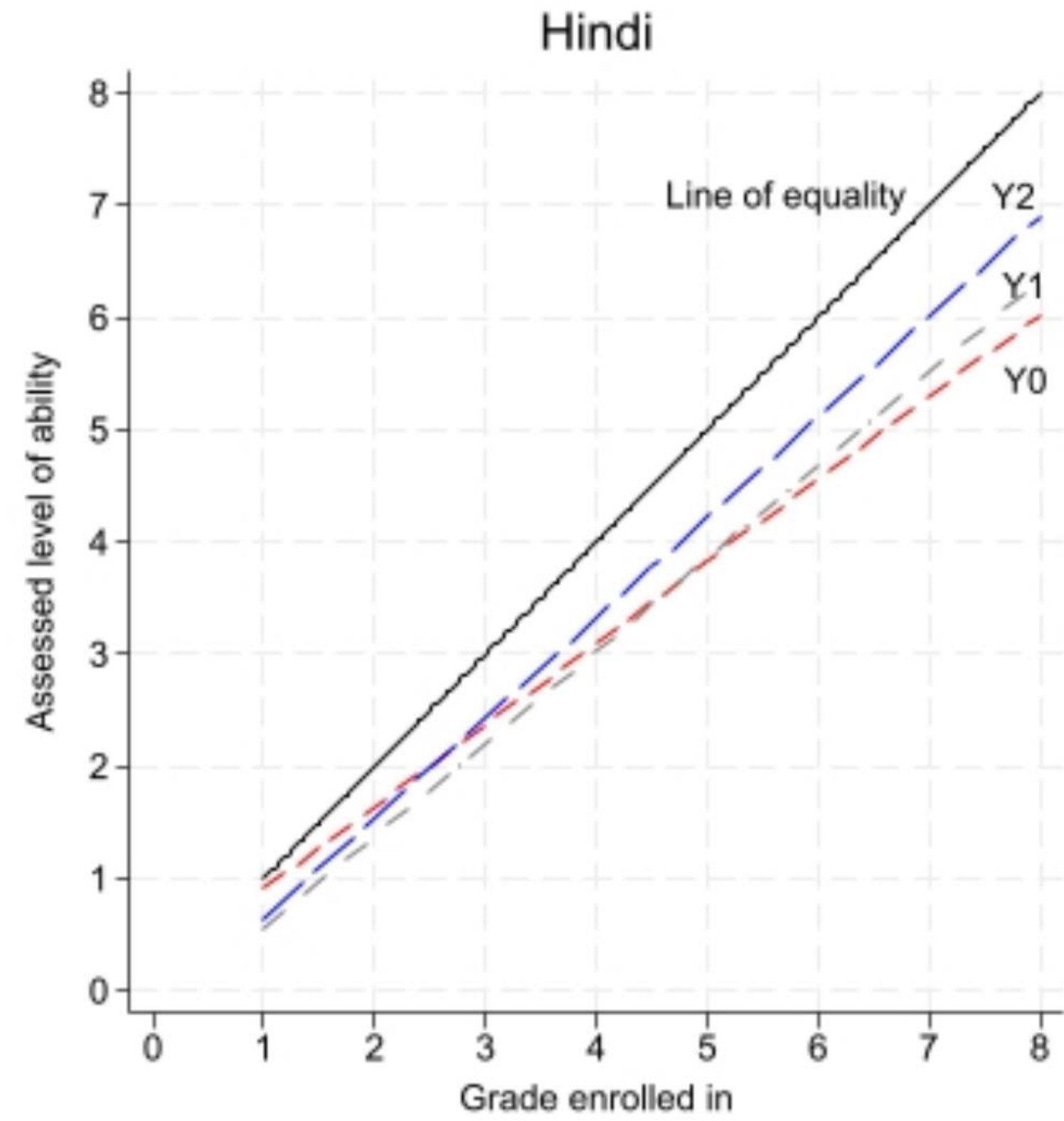
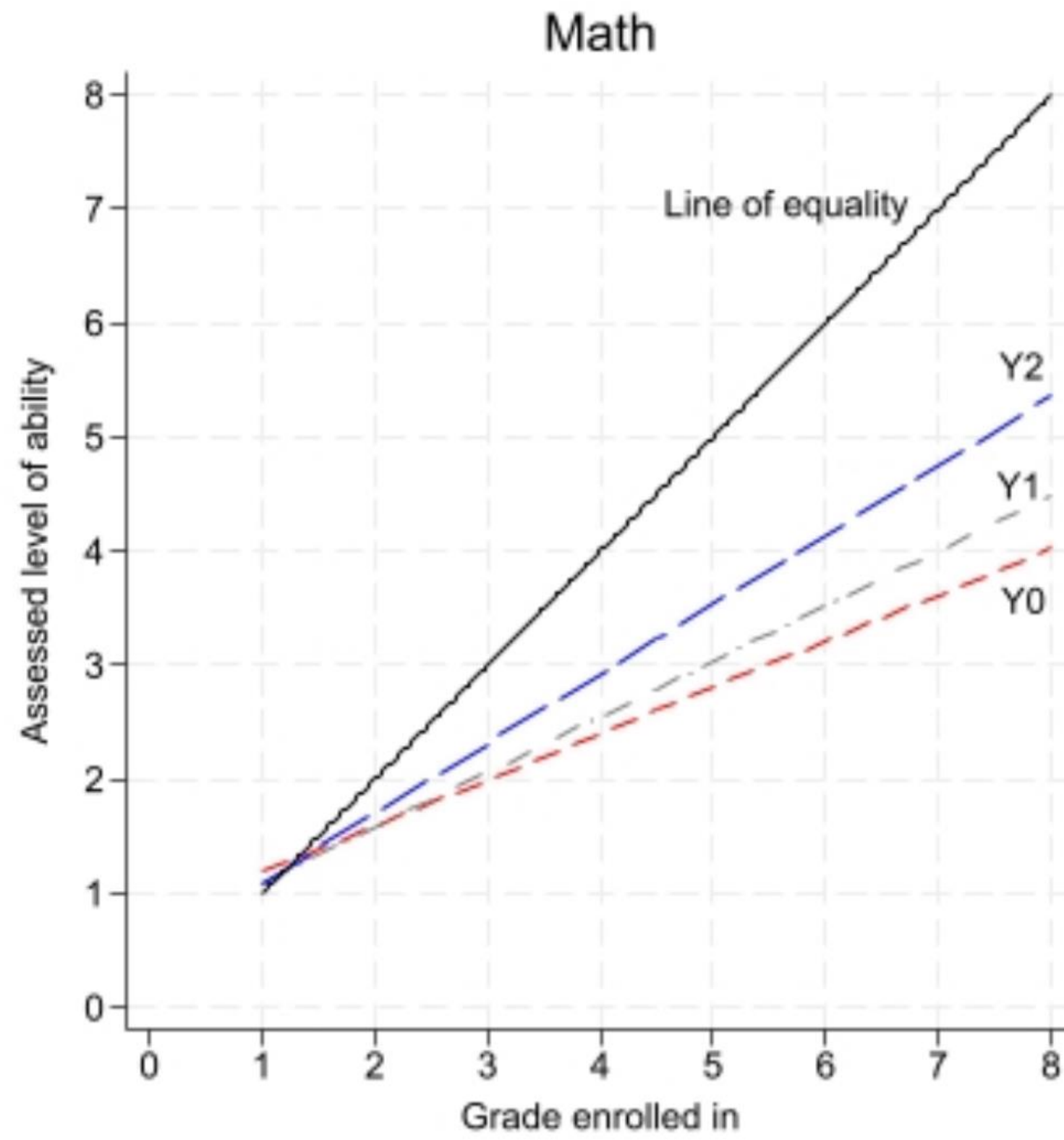
Effects after 18 months of treatment



Interpreting magnitudes

- ▶ Program effects equal 48% of business-as-usual value-added in math, 65% in Hindi, in the control group over 18-months
 - ▶ Estimated using vertically-linked IRT scores
 - ▶ These can be interpreted as productivity gains (delivered in the school day)
- ▶ Same range as other promising in-school interventions:
 - ▶ **“Teaching at the Right Level”:**
 - ▶ Balsakhi program ([Banerjee et al 2007](#)): 0.28σ
 - ▶ Tracking in Kenya ([Duflo, Dupas and Kremer, 2011](#)): $0.14\text{--}0.19\sigma$
 - ▶ **Incentives in schools:**
 - ▶ Performance bonus ([Muralidharan and Sundararaman, 2011](#)): $0.16\text{--}0.22\sigma$
 - ▶ Cameras in schools ([Duflo, Hanna and Ryan 2012](#)): 0.17σ
- ▶ This is probably easier to scale than incentives, in an area that governments do want to invest in

The program reduced the gap between student ability and the curriculum



--- Year 0 - - - Year 1 - - - Year 2 — Line of equality

Insights from further adaptation for scale in Year 3

- ▶ **In Y3, the intervention reduced in-person support for implementation**
 - ▶ From one lab-in-charge per school to 1:3
- ▶ **Usage sharply reduced in the short run**
 - ▶ ~50% lower in July Y3 compared to July Y2
 - ▶ Recovered with program refinement by Oct (since usage is monitored real-time)
- ▶ **Program value-added declined but dose response did not**
 - ▶ The instruction did not become less productive, there was just less of it
- ▶ **Highlights that the usage data in EdTech platforms is both a metric for take-up *and* a metric for program refinement**

Discussion

- ▶ **We solve a substantive policy problem:** how to integrate EdTech into classrooms and school schedules without adding instructional time
 - ▶ Our results show that these can feasibly lead to large productivity gains
 - ▶ Cost-effectiveness 2x of original efficacy trial
- ▶ **The problems we solve are likely to be relevant in many global settings:**
 - ▶ Slow progress, substantial dispersion within a classroom
 - ▶ The perceived promise of EdTech is likewise global
- ▶ **Our results also demonstrate the value of staged innovation**
 - ▶ Underlies e.g. the approach of USAID DIV, Global Innovation Fund, AFD FID
 - ▶ This iteration was funded by the Global Innovation Fund
 - ▶ Underemphasized in the academic literature
- ▶ **Highlights a distinction, even within development, about scaling up from efficacy trials:**
 - ▶ The best evidence on scaling comes from programs involving a discrete input directly to end users through a purpose-built delivery chain (deworming, cash transfers, microfinance...)
 - ▶ This is very different from scaling up processes (like daily instruction) within organizations
 - ▶ Requires system-level buy-in and constant effort by implementing agents (here, staff and students)

Thank you

- ▶ Questions? Thoughts? Comments?
- ▶ Please reach out: kamurali@ucsd.edu | abhijeet.singh@hhs.se

Bibliography I

- Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131(3), 1415–1453.
- Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., ... Udry, C. (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236), 1260799.
- Banerjee, A., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1), 1–21.
- DellaVigna, S., & Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1), 81–116.
- Kraft, M. A., Schueler, B. E., & Falken, G. (2024). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability. edworkingpaper no. 24-1031. *Annenberg Institute for School Reform at Brown University*.
- Linós, E., Prohofskey, A., Ramesh, A., Rothstein, J., & Unrath, M. (2022). Can nudges increase take-up of the eitc? evidence from multiple field experiments. *American Economic Journal: Economic Policy*, 14(4), 432–452.
- List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Currency.

Bibliography II

- List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature*, 626(7999), 491–499.
- Mitchell, H., Mobarak, A. M., Naguib, K., Reimão, M. E., & Shenoy, A. (2023). Delegation risk and implementation at scale: Evidence from a migration loan program in bangladesh. *Unpublished manuscript*. https://ashishenoy.github.io/Website/Paper_NLS_Evaluation.pdf.
- Muralidharan, K., & Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, 31(4), 103–24.
- Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2016). Building state capacity: Evidence from biometric smartcards in india. *American Economic Review*, 106(10), 2895–2929.
- Muralidharan, K., & Singh, A. (2020). *Improving public sector management at scale? experimental evidence on school governance india* (Tech. Rep.). National Bureau of Economic Research, Inc.
- Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4), 1426–60.