

High Fidelity, Uneven Impacts

When Observed Program Delivery and Mechanism Activation Diverge

Fernanda Soares · TJ D'Agostino · Hannah Reynolds · Patrizio Piraino ·

In partnership with Associate in Research and Education for Development (ARED)

The scaling problem

THE PROBLEM

"Voltage Drop"

Educational interventions that work in pilots regularly fail to replicate at scale, even when implementation looks faithful.

THE GAP

Fidelity $\neq$ Effectiveness

Existing frameworks measure whether activities were delivered as planned. They rarely ask whether the mechanisms for change were actually activated.

OUR ARGUMENT

A Diagnostic Approach

Within-trial heterogeneity is not noise, it is signal. Variation across sites may reveal what conditions turn delivery into learning.

Program & Research Design

Ndaw Wune Vacances (NWW)

An eight-week summer remedial intervention combining structured pedagogy and targeted instruction based on students' learning levels | Grades 1–2 | ARED, Senegal



Cluster-Randomized Controlled Trial

100 schools (50 treatment, 50 control), commune-stratified.



Implementation Monitoring

Fidelity of implementation, attendance, coaching frequency, tutor characteristics



Qualitative Inquiry

Purposive sample of 15 schools (5 high / 5 medium / 5 low impact). In-depth interviews and visit-log records

Diagnostic account of within-trial heterogeneity & prospective voltage drop

The program worked, on average

Intent-to-Treat (ITT) effects are positive and statistically significant for both literacy and math outcomes

+0.23***

SD

LITERACY

Familiar Words +0.18*** SD

Letter Knowledge +0.13 SD

Oral Reading Fluency +0.20*** SD

Reading Comp ORF +0.24** SD

Reading Comp Pic Matching +0.23** SD

Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

+0.15**

SD

MATHEMATICS

Number Identification +0.10** SD

Counting +0.11 SD

Number Discrimination +0.11 SD

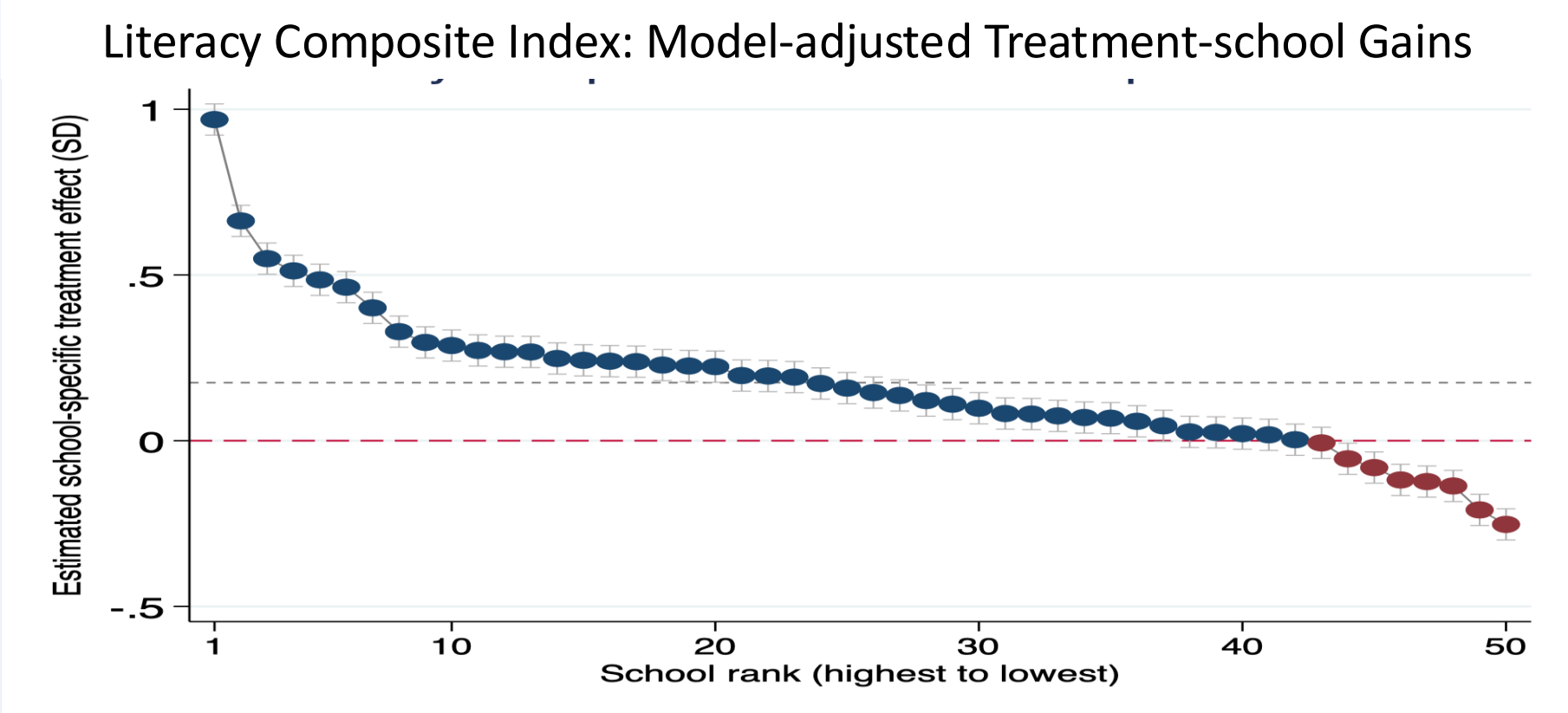
Addition +0.13* SD

Subtraction +0.11 SD

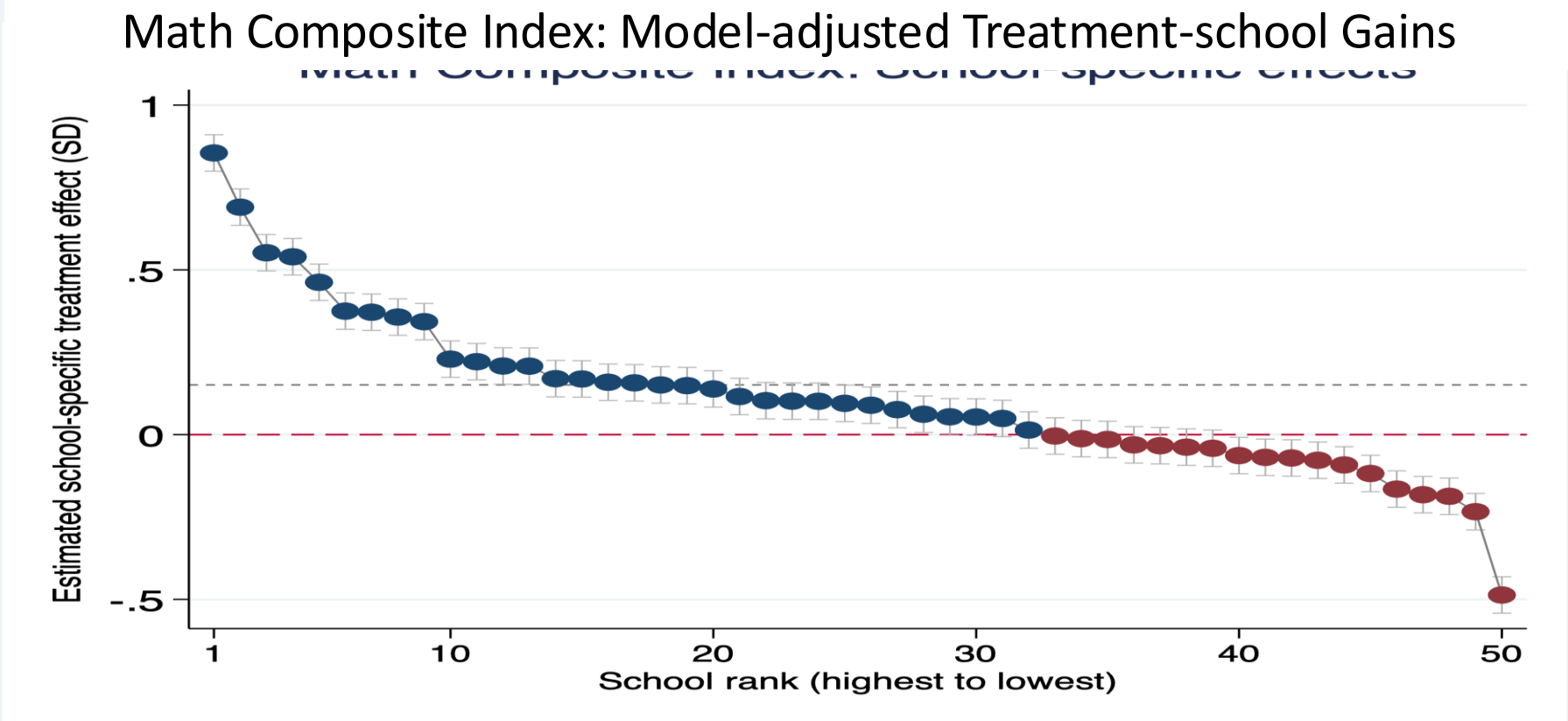
Note. * $p < .05$. ** $p < .01$. *** $p < .001$.

Heterogeneity across schools

The average conceals a striking range: most gains are concentrated in a small number of schools



87.6% of total positive literacy impact concentrated in top half of schools

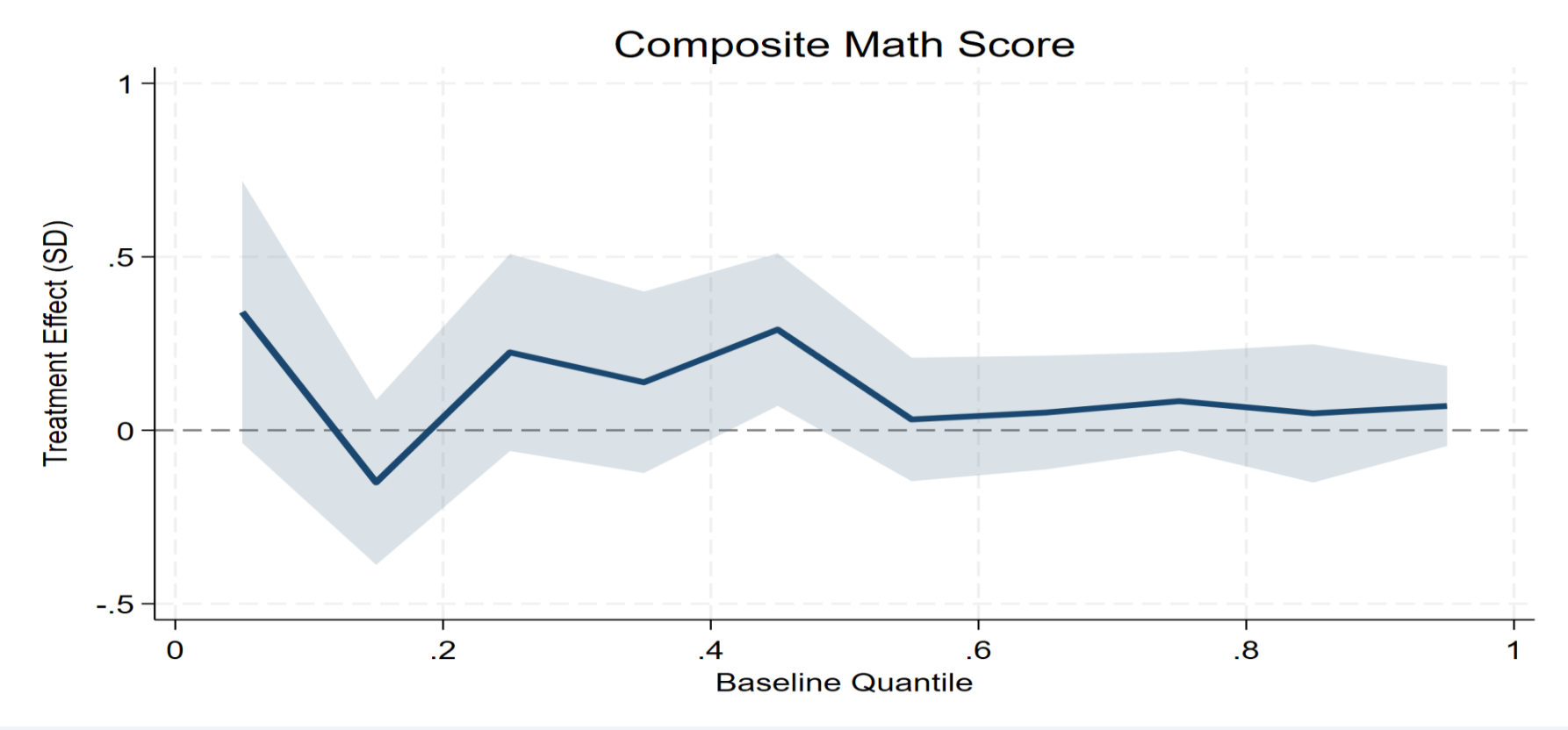
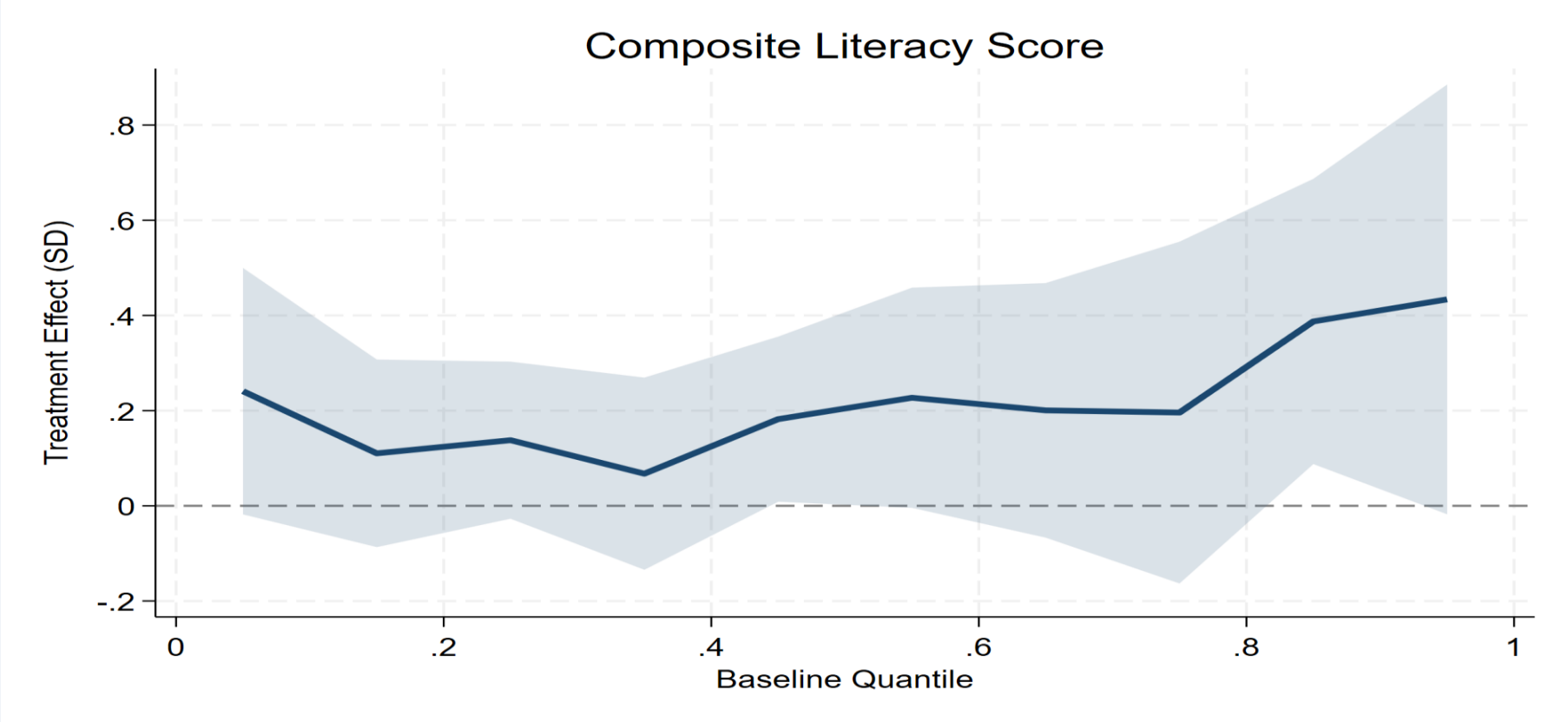


94.7% of total positive math impact concentrated in top half of schools

Treatment schools ranked by model-adjusted treatment-school gains relative to the pooled control trend. Red dashed = zero; grey dashed = average ITT. 95% CIs with school-clustered SEs. Interpret as broad indicators, not precise rankings.

Heterogeneity across learners

Effects are positive across most deciles in literacy, with a more complex and uneven pattern in mathematics



LITERACY

Gains are largely positive and increasing toward higher baseline levels, indicating that higher-skilled students benefited more.

MATH

More complex and uneven: some lower-decile gains exceed middle-decile ones; effects are slightly larger at lower baseline levels and flatten at higher quantiles.

Estimated treatment effects by decile of baseline achievement using quantile regression. Shaded areas = 95% CIs.

What explains the variation?

✓ ATTENDANCE

Higher attendance was associated with stronger learning gains, especially in literacy. Math effects were smaller but generally positive.

✗ FIDELITY OF IMPLEMENTATION (DELIVERY)

Fidelity to the NWV teaching approach - sequencing, level-based grouping, material use, autonomous-group instructions and tutor rotation, - was uniformly high. Yet fidelity was only weakly related to school-level gains.

Other implementation features and tutor characteristics:

Coaching frequency, supervisor use of assessment data, tutor characteristics — prior programme experience, teaching degree, years of experience — do not meaningfully explain why program impacts differed so markedly across schools.

What conditions enabled mechanism activation?

01

KEY MECHANISM

Supervisor Stewardship

High-impact schools had “NWV stewards” who reinforced the methodology, tracked student progress, protected sequencing, and managed disruptions.

Medium- and low-impact schools had more generic or weakly documented supervision.

02

KEY MECHANISM

Managing Disruptions Adaptively

Every school faced the same disruptions: rain, farm labor, religious events.

High-impact schools organised make-up sessions, did proactive parent outreach, and individual follow-up.

Lower-impact schools accommodated disruptions without recovering lost learning time.

03

KEY MECHANISM

Protecting Instructional Time

High-impact schools used flexibility to preserve lesson continuity and content coverage.

Lower-impact schools allowed disruptions and adaptations to erode the instructional logic of the program.

High-impact schools did not simply deliver NWV; they actively protected the conditions needed for the program’s learning mechanisms to work.

Why do lower-baseline students benefit less?

The program targets low-baseline students, yet gains were more reliable among students who entered with stronger foundational skills

"When you put a struggling student in an independent group, it's pointless."

— NWV Tutor, qualitative interview

✓ For higher-baseline students

Foundational skills allow students to work productively in autonomous groups, even with limited tutor attention. They can learn from peers and materials.

Independent time becomes learning time.

✗ For lower-baseline students

Students struggle to self-manage the task. Without tutor support, independent time becomes idle time or students drift back toward the direct instruction group.

The independent learning mechanism is not activated.

What this study tells the field

Delivery-activation decoupling: Programs can follow observable protocols without activating the instructional, relational, and organizational mechanisms needed to generate learning.

THEORETICAL CONTRIBUTIONS

01 Mechanism activation $\neq$ fidelity to protocols

High fidelity to protocols does not guarantee learning mechanisms are activated.

02 Site-level heterogeneity is signal, not noise

Variation across schools reveals where and why mechanisms work.

A CALL TO ACTION

01 Study cross-site heterogeneity

Power and design studies to examine variation across schools, not only average effects.

02 Expand measurement beyond fidelity

Capture implementation quality, participant responsiveness, and mechanism activation, not only protocol adherence.

Scaling implication: Scaling requires systems that help learning mechanisms activate reliably across diverse schools, students, and implementers.

Thank you!

Fernanda Soares (fernanda.soares@nd.edu)

"Behavioural science is unlikely to change the world without a heterogeneity revolution"
Bryan, Tipton & Yeager (2021, Nature Human Behaviour)