

The Illusion of Comparability Among Standardised Effect Sizes: Why Education Evaluations Should Report Raw Effects

Jack Rossiter, David Evans, Susannah Hares, Catherine Henny

WWHGE Conference
16 September 2026

Motivation.

What are children actually learning
to do?

What it said...

“...overall reading improves by 0.64 SDs... and overall writing performance by 0.45 SDs. These reading and writing effects are comparable to some of the largest measured in the literature.”

“... literacy and numeracy skills increased by 0.46 and 0.63 standard deviations... the combined intervention is highly effective and cost-effective.”

“After 2.5 years, the program increased... test scores by 0.41 standard deviations.”

What it meant...

0.64 SDs

+10 more letters sounded
+2 more words per minute
+0.2 reading comp. answers

0.46 SDs

+ 1.5 more letters sounded
+ 0.3 more words per minute
+ 0.4 listening comp. answers
~ 0 change in reading comp.

0.41 SDs

+1 additional letter
or word identified,
from eight

Identify this letter		Read this word				Identify the missing word	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
C	T	PAPA	VÉLO	ÉCOLE	TOMATE	Le garçon achète un _____. A. HEUREUX B. BONBON C. EST	La fille ____ aller à l'école. A. JOUR B. CHEMIN C. AIME

What we did

Analyse the relationship between standardised effects & skills changes across a broad literature

Process: build out from four large evidence reviews (Evans & Yuan, 2022; GEEAP, 2023; Kim et al., 2020; Bertling et al., 2023) + update search, to identify:

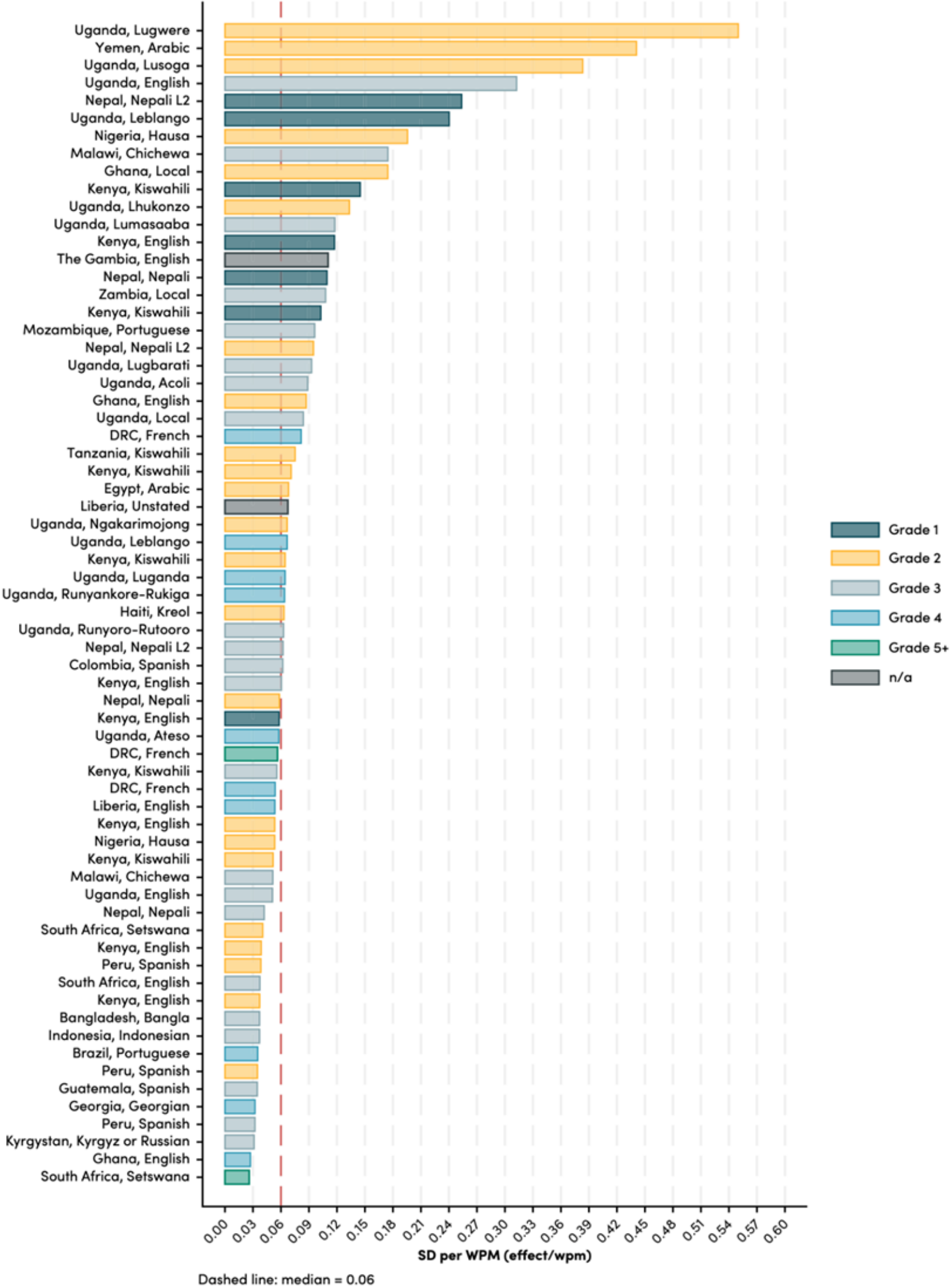
- peer-reviewed papers that;
- are conducted in a low or middle income country;
- and measure learning effects for primary-school children;
- and report at least one reading, language or literacy outcome;
- and use an experimental or quasi-experimental design.

Yields: 171 papers containing 2,018 reading or language effects.

The exchange rate we use varies a lot

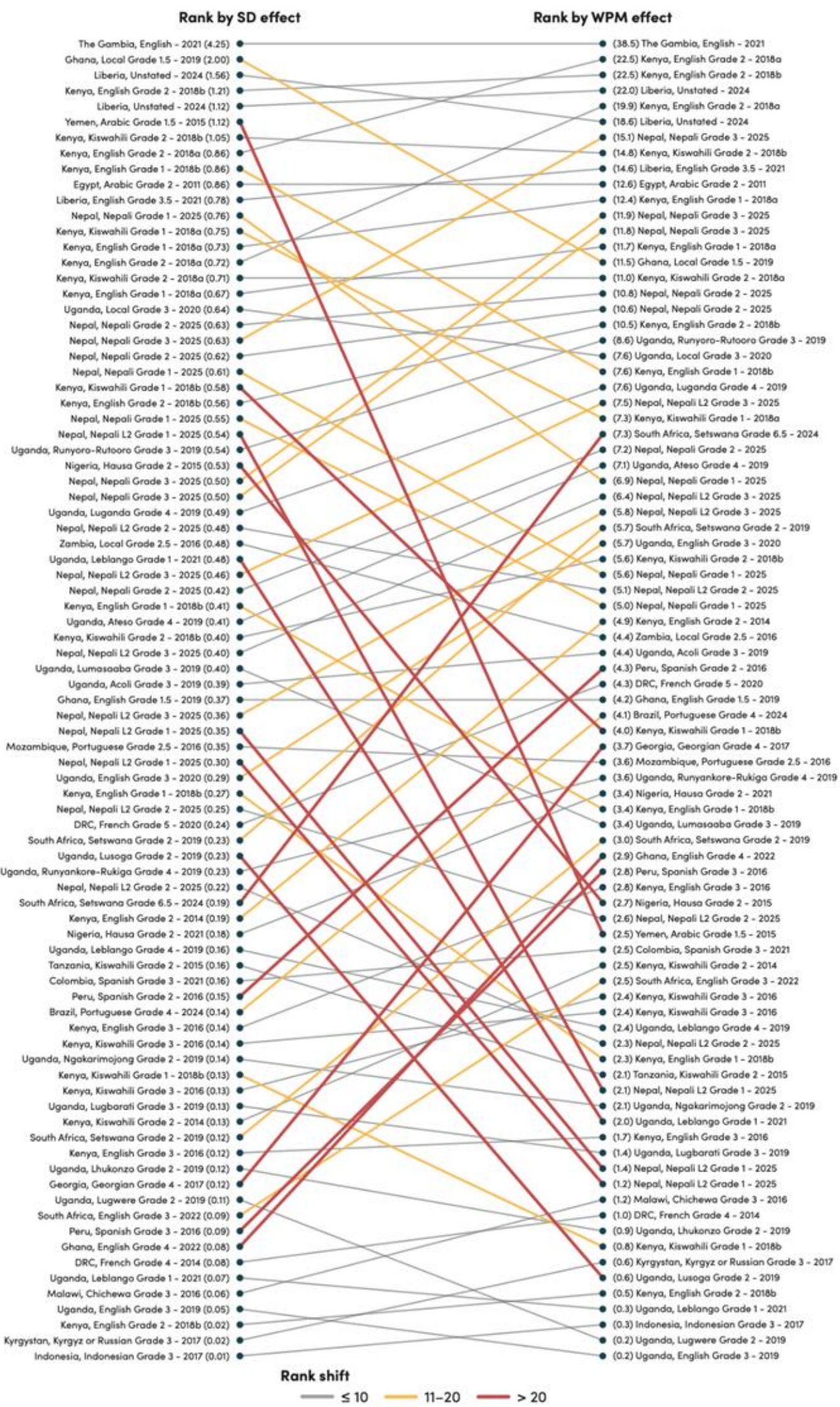
“Exchange rate” = standard deviations per word-per-minute gained

Even accounting for language or grade differences, the impact of a one-word improvement cannot plausibly differ by factors of **six** or **twenty**.



The consequences of this extend to the ranking of programmes

- Despite strong *overall agreement*, individual observations shift substantially.
- The largest movers are interventions in lower literacy contexts where small absolute reading fluency gains produce large standardised effects.



Many other factors influence the reported effect

How Learning and Skills are Measured	How Tests are Scored, Standardised and Reported	How Intervention Design and Analysis are Specified
What construct and skill domain does the instrument cover?	What is the score distribution used to standardise the raw effect?	Is the evaluation of a causal or a correlational relationship?
How well was the measure aligned to the abilities of students tested?	Are results reported overall or by subgroup?	What is the sample size of students and schools/sites?
In what language was the assessment conducted?	How are tasks or test items scored and combined?	What does the comparison group experience while the intervention takes place?
Does the assessment meet basic psychometric standards?	Where there are subtasks, which are reported and which are highlighted?	Over what time period were impacts measured?

Response.

Ask: “What is it that children can do now, thanks to the programme?”

Three steps towards better reporting

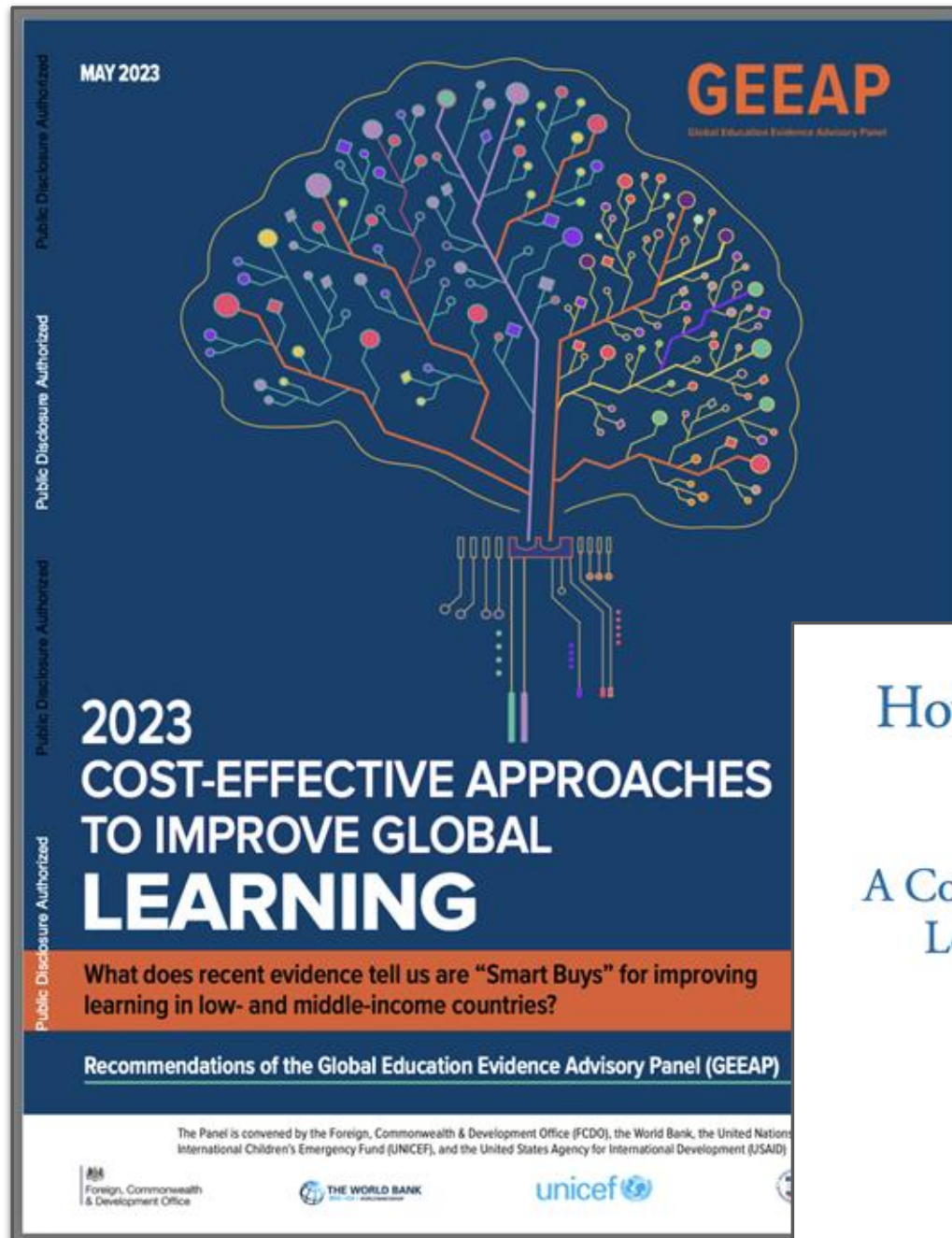
- Step 1. Greater transparency
- Step 2. Use benchmarks to put skill changes in perspective
- Step 3. Use measures that are fit for the skills being assessed

What we are doing next and where we need help

- Consult, develop and test a short reporting guide aimed at **authors and reviewers**
- Convene **funders** to think through minimum standards for reporting learning effects & get organisations to commit to that
- Communicate the main message for **evidence users** e.g. an infographic to make the problem immediately clear

Additional content

We increasingly want to (and can) compare programs



Briefs

Educational Evaluation and Policy Analysis

September 2022, Vol. 44, No. 3, pp. 532–540

DOI: 10.3102/01623737221079646

Article reuse guidelines: [sagepub.com/journals-permissions](https://journals.sagepub.com/journals-permissions)

© 2022 The Author(s). <https://journals.sagepub.com/home/epa>



How Big Are Effect Sizes in International Education Studies?

David K. Evans

Center for Global Development

Fei Yuan

Harvard Graduate School of Education

How to Improve Education Outcomes Most Efficiently?

A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric

Noam Angrist

David K. Evans

Deon Filmer

Rachel Glennerster

F. Halsey Rogers

Shwetlena Sabarwal

One question we want to answer from these comparisons:



Briefs

Educational Evaluation and Policy Analysis

September 2022, Vol. 44, No. 3, pp. 532–540

DOI: 10.3102/01623737221079646

Article reuse guidelines: sagepub.com/journals-permissions

© 2022 The Author(s). <https://journals.sagepub.com/home/epa>



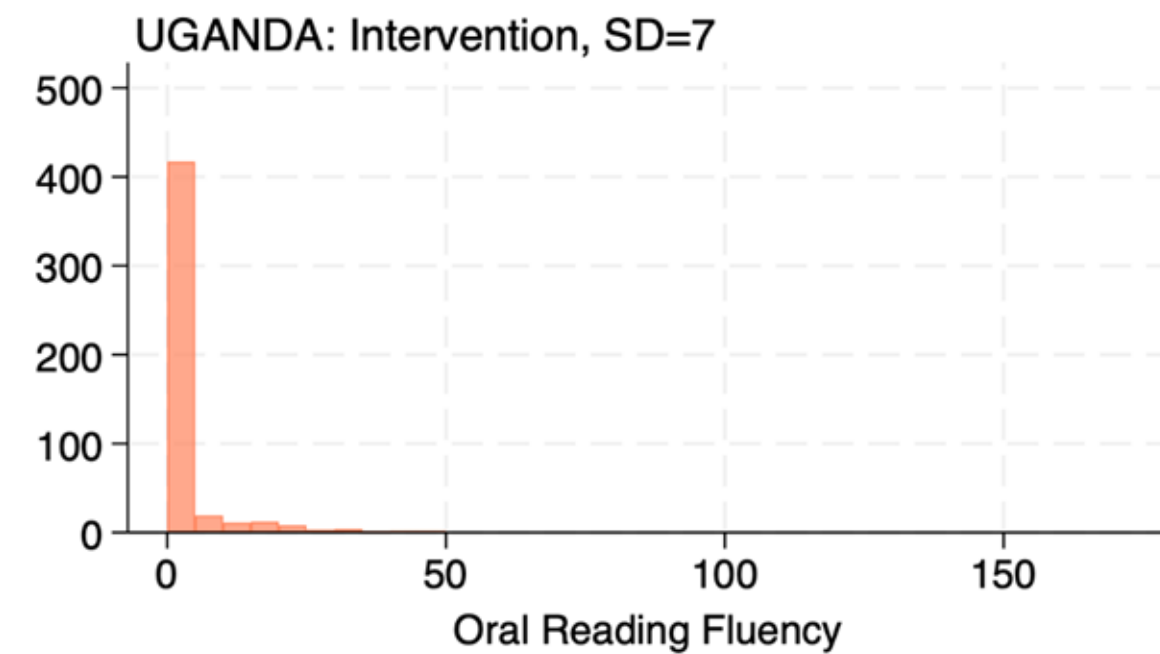
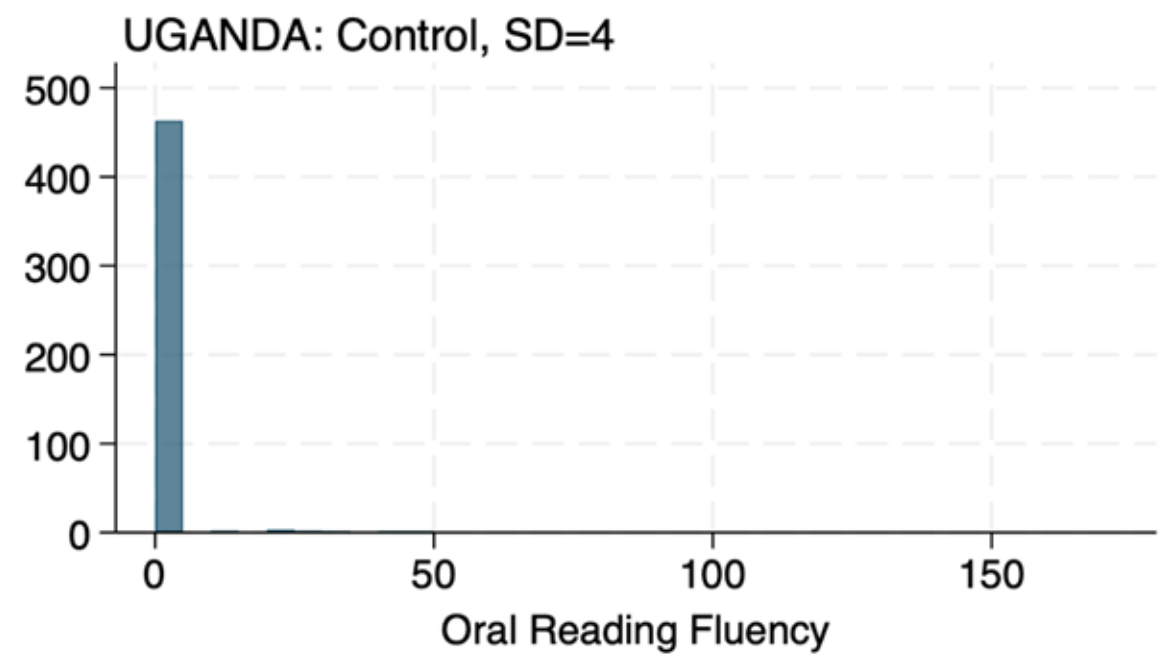
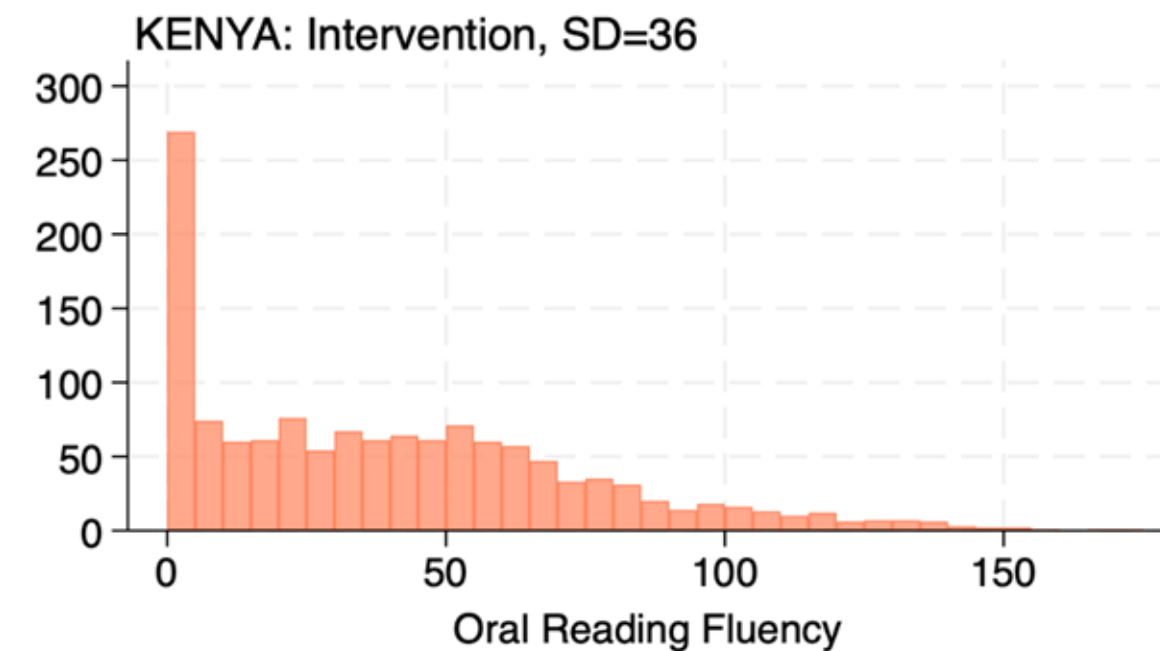
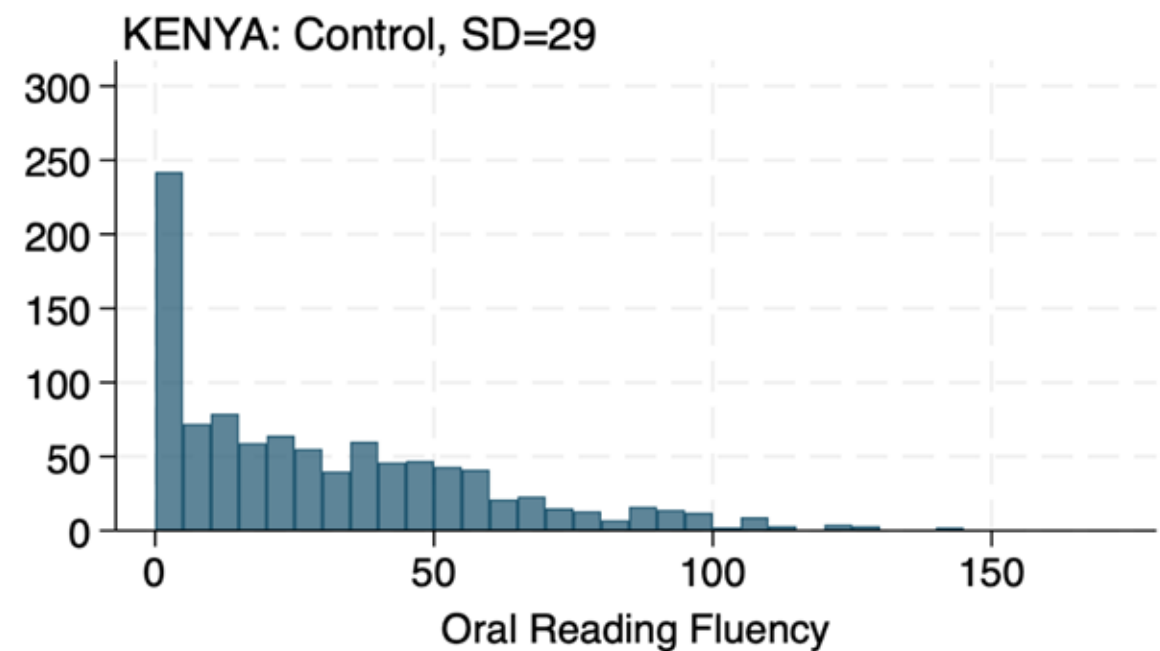
How Big Are Effect Sizes in International Education

Which approaches move children furthest
towards reading with understanding?

*A Comparison of 150 Interventions Using the New
Learning-Adjusted Years of Schooling Metric*

*Noam Angrist
David K. Evans
Deon Filmer
Rachel Glennerster
F. Halsey Rogers
Shwetlena Sabarwal*

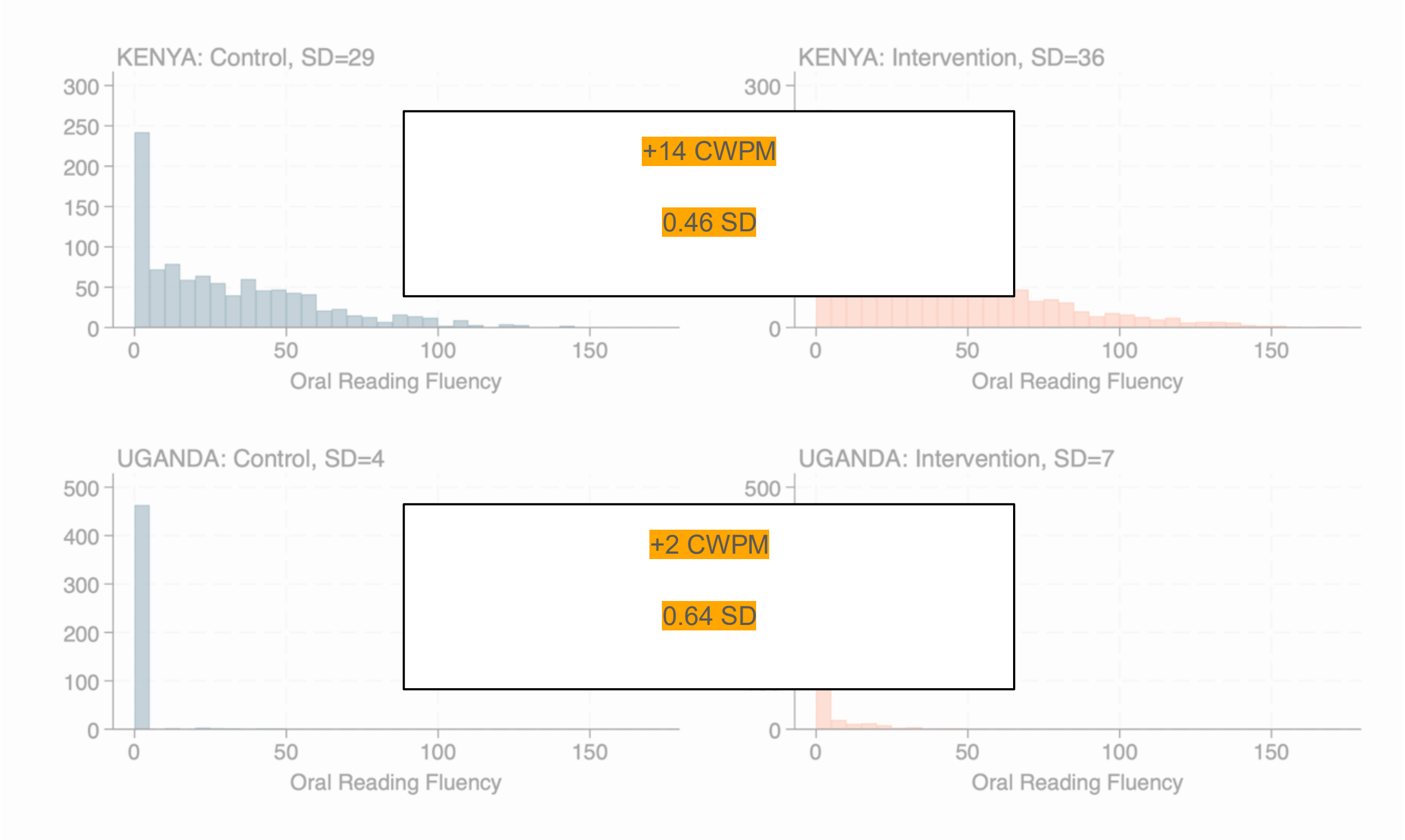
When almost nobody can read, a tiny gain *can* look enormous



When almost nobody can read, a tiny gain *can* look enormous

In Uganda, only 21 of 473 children assessed could read a single word.

There is effectively no ‘distribution’ of reading fluency to talk about.



It’s not that hard & puts constructive pressure on researchers

“It is useful to pause and consider how large these effects are... [At baseline] 39 percent of the children could not recognize letters, and only 15 percent could read a paragraph or story... At endline, there was little progress in the control group... But in the treatment group, only 8 percent could not recognize letters, and 49 percent could read a paragraph or a story.”

TABLE 3—TREATMENT EFFECT BY SPECIFIC COMPETENCE ASSESSED

	Proportion of questions answered correctly			
	Sentence completion	Retrieve explicitly stated information	Make straight-forward inferences	Interpret and integrate ideas and information
<i>Panel B. Hindi</i>				
Treatment	0.046 (0.022)	0.045 (0.016)	0.065 (0.022)	0.053 (0.015)
Baseline Hindi score	0.13 (0.017)	0.14 (0.0075)	0.15 (0.011)	0.067 (0.013)
Constant	0.72 (0.011)	0.59 (0.0078)	0.51 (0.011)	0.31 (0.0077)
Observations	539	539	539	539
R ²	0.182	0.380	0.309	0.136

Three steps support progress towards reading (effects) with meaning

Key question: “What is it that children can now do, thanks to the program?”

Step 1. Greater transparency

- Report raw gains, ideally in clear skill terms
- Report details of reference group and robustness

Three steps support progress towards reading (effects) with meaning

Key question: “What is it that children can now do, thanks to the program?”

Step 1. Greater transparency

Step 2. Use benchmarks to put skill changes in perspective

- Units with clear meanings
- Anchor results to specific, age-appropriate skills or competencies

Three steps support progress towards reading (effects) with meaning

Key question: “What is it that children can now do, thanks to the program?”

Step 1. Greater transparency

Step 2. Use benchmarks to put skill changes in perspective

Step 3. Use measures that are fit for the skills being assessed

- Document the instrument
- Develop publicly available item banks (GEEAP/LAYS)