CENTER FOR GLOBAL DEVELOPMENT

The Illusion of Comparability Among Standardised Effect Sizes

WHY EDUCATION EVALUATIONS SHOULD REPORT RAW EFFECTS

Jack Rossiter, David K. Evans, Susannah Hares, and Catherine Henny

Abstract

Policymakers, practitioners, and donors rely on standardised effect sizes to compare education programmes and allocate resources. Yet in early-grade reading interventions in low- and middle-income countries, these metrics reflect differences in measurement tools, sample composition, and scaling choices as much as genuine learning gains. Drawing on data from 197 studies, we find that a single additional word per minute of reading fluency corresponds to anywhere from 0.03 to 0.55 standard deviations across studies. For tests designed by researchers—the most common type of test in this literature and the hardest to compare—a single correct response on a reading assessment can result in a standardised gain anywhere from 0.08 to 0.80 standard deviations. Converting raw to standardised effects also reorders which programmes appear most effective. Fewer than one in four papers (23 percent) present both standardised and raw effects, leaving readers unable to judge specifically what children can do differently as a result of the programme and whether that change matters educationally. We propose that researchers report raw effects and reference distributions alongside standardised estimates, that work on reading benchmarks be extended across more languages and countries, and that shared measurement frameworks be developed faster. Together, these reporting changes can improve interpretation for policymakers and allow evidence syntheses to explicitly model differences in test design and sample composition across studies. Each of these changes would help the field show what children actually gained from a programme, not just how large the effect appears.

WORKING PAPER 748 • JUNE 2026

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

Jack Rossiter
Center for Global Development (non-resident fellow)

David K. Evans
Center for Global Development

Susannah Hares
Independent

Catherine Henny
Independent consultant

Corresponding author: Rossiter (jack.rossiter@barcelonagse.eu). Other authors are listed alphabetically. This study was financed by the Luminos Fund. Thank you to Suraya Alami and Sébastien Hine for research assistance. Thank you to Noam Angrist, Caitlin Baron, Matthew Jukes, Michelle Kaffenberger, and Kirsty Newman for constructive comments. The authors are responsible for any remaining errors.

Replication data and code are available at https://doi.org/10.7910/DVN/TWMP8S

Jack Rossiter, David K. Evans, Susannah Hares, and Catherine Henny. 2026. "The Illusion of Comparability Among Standardised Effect Sizes: Why Education Evaluations Should Report Raw Effects." CGD Working Paper 748. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/illusion-comparability-among-standardised-effect-sizes-why-education-evaluations-should

CENTER FOR GLOBAL DEVELOPMENT
2055 L Street, NW Fifth Floor
Washington, DC 20036

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

Contents

Figures

Tables

1. Introduction

Policymakers, practitioners, and donors want to know which education interventions lead to meaningful improvements in children's reading and language skills in low and middle income countries.[1] Standardised effect sizes are the dominant metric for summarising and comparing those improvements across interventions (Evans & Yuan, 2022; Singh, 2015a; Angrist et al., 2025). They provide a common scale when studies use different instruments, populations, and designs, and they let readers gauge gains even across unfamiliar tests. These metrics form the foundation of the evidence synthesis infrastructure on which resource allocation decisions increasingly depend (Ritchie, 2025).

The problem is not that researchers standardise, but that standardised effects are routinely reported in isolation, without the contextual information needed to judge what they mean. Variation in reported effects reflects differences in measurement tools, sample characteristics, and scaling choices as much as underlying learning gains. The pursuit of larger standardised effects can therefore be misleading unless we understand what they mean in terms of raw, interpretable changes in children's reading ability. A simple example illustrates how far this variation can go. A one-word gain in oral reading fluency can have a six-times larger standardised effect size in a study from Uganda (Leblango) than one in Kenya (English), and is more than twice the size in a study in The Gambia (English). In Nigeria (Hausa), a single additional word is almost twenty times the value of a word in South Africa (Setswana). Even accounting for language or grade differences, the substantive impact of a one-word improvement cannot plausibly differ by factors of six or twenty.

Such large differentials become a major concern in a literature that has grown rapidly and is used frequently for cross-study comparison, ranking, and policy guidance. In 2006, a review of randomised controlled trials (RCTs) of school-based interventions with learning outcomes could cite only seven studies from three countries (Glewwe & Kremer, 2006).[2] A decade later, the landscape had transformed: a similar review documented 118 high-quality studies, including 80 RCTs (Glewwe & Muralidharan, 2016). There have been at least ten more reviews published since then.[3] This growth reflects both an expansion of studies with credible causal claims and a strong motivation to generate and synthesise evidence to identify the most effective ways to allocate scarce education resources.

Yet progress in measuring skills outcomes within these evaluations has not kept pace with the growth in the production of causal evidence. As the literature has expanded, so too has the diversity of learning assessments, intervention types, and target populations being compared.

1 It is not the objective of this paper to provide a definitive judgment as to what constitutes a meaningful improvement. We would propose that a one-word gain in oral reading fluency is not a meaningful improvement and that learning to read a paragraph with comprehension is a meaningful improvement. Between those extremes, opinions on what improvement is meaningful may vary dramatically.

2 Radio mathematics programmes, or mathematics workbooks in classrooms (Nicaragua), Uniforms, textbooks and classroom buildings (Kenya), Textbooks (Kenya), Deworming medicine (Kenya), Flipcharts (Kenya), Community teachers (India), Computer assisted learning (India).

3 Ganimian & Murnane (2016), Conn (2017), J-PAL (2017), Asim et al. (2017), Graham & Kelly (2019), Kim et al. (2020), GEEAP (2020; 2023b), Evans & Yuan (2022), Angrist et al. (2025).

Glewwe & Muralidharan (2016) warned that "more public goods and standards for measurement and reporting need to be created to make it easier for highly decentralized (and often opportunistically conducted) research studies to be compared across contexts." Despite repeated calls for better, directly comparable outcome metrics and acknowledgements that standard deviations should be benchmarked against real world measures, little has changed (Bertling et al., 2023; Evans & Yuan, 2022; Stern & Piper, 2019). Meta-analyses often rest on the assumption that larger effect sizes imply greater educational significance and can be meaningfully combined across studies, an assumption that is unsafe when underlying study details differ (Simpson, 2017).

On average across papers, reporting practices lack transparency. We document that unstandardised ("raw") effect sizes are rarely reported alongside their standardised counterparts. Concepts such as Learning Adjusted Years of Schooling (which currently take standardised effects as inputs) rescale effects against a further standard deviation, potentially compounding any distortion in the original estimate.[4] Even within-study translations, for example expressing a programme's impact as equivalent to a given number of additional months of schooling, rely on a learning rate drawn from that study's same reference group with dispersion that is highly variable across studies. The further such conversions move from raw outcomes, the wider the disconnect between what is reported and what children can actually do (Kraft, 2020; Baird & Pane, 2019), and the harder it becomes to communicate the meaning of programme effects to the practitioners and policymakers who must act on them (Cuijpers, 2021). The risk is that resource allocation decisions may reflect statistical artefacts as much as genuine improvements in children's skills.

Drawing on data from 197 studies of education interventions in low and middle income countries, this paper asks how far standardised effects can be compared across studies when the underlying measures, score distributions, and standardisation choices differ. Using oral reading fluency outcomes, we show that access to unstandardised effects and basic contextual information can materially change judgments about which programmes may merit scale-up or adaptation into new contexts. We then demonstrate how dispersion in the reference group, the approach to standardisation, and the suitability of the measurement instrument can produce large apparent differences in programme effectiveness even when underlying learning gains are similar. These issues are connected (instrument choice can shape dispersion, and dispersion shapes the standardised effect), and so are the three changes we propose in response, ordered from the most immediately implementable to the most ambitious.The most immediate step is greater transparency in reporting. We propose that researchers report two easily derived pieces of information alongside any standardised effect. First, the unstandardised ("raw") effects, expressed in terms that show what children can now do as a result of their participation. Second, the reference sample distribution used to calculate the standard deviation when standardising. Reporting these two pieces of information would also place constructive pressure on researchers to engage more directly with the properties

4 Learning Adjusted Years of Schooling could accept any comparable learning unit, such as words per minute (Angrist et al., 2025). Standardised effects are used because of their current availability and flexibility across studies, but work to develop directly comparable scales may reduce this dependence.

of the assessments they use, including their content, structure and reliability, the suitability of the chosen outcome for the sample, and whether the observed gain represents a meaningful shift in children's skills. We don't argue for abandoning standardisation but rather for routinely supplementing it with basic contextual information that researchers already collect.

Two further changes will take longer but should be accelerated. Outcomes need to be anchored to simple, language appropriate benchmarks (e.g., Ardington et al., 2021; Spaull et al., 2020) so that gains can be read as movements towards skills that matter. And the field needs faster progress on shared measurement frameworks because the choice of instrument is fundamental to whether a standardised effect can carry comparable meaning across studies at all. Both are important for connecting claims about programme effectiveness to genuine improvements in skills and for making evidence synthesis more informative for policy.

Several authors have examined the relationship between standardised effects and underlying learning in general terms (Alsalti et al., 2024; Baguley, 2009; Cuijpers, 2021; Kraft, 2020; Simpson, 2017), and others have raised measurement challenges specific to developing country contexts (Singh, 2015a; Cheung & Slavin, 2016; Bertling et al., 2023). Some have also demonstrated the problem empirically. Stern & Piper (2019) showed across six interventions that large effect sizes frequently failed to translate into meaningful gains in the proportion of children reading at benchmark. Their purpose was to help programme designers set realistic targets, but the finding speaks directly to the reliability of effect sizes as a summary of impact. Kerwin & Thornton (2021) show that, within a single randomised trial, the apparent programme effect is highly sensitive to which literacy subtask is used as the outcome. And Singh (2015b) notes that applying different scoring methods to the same set of test responses can double the estimated effect. These studies raise an important alarm, with each focused on a specific mechanism within a small set of programmes or a single evaluation. This paper attempts to describe how these and other distortions operate across the broader evidence base, discussing what that means for cross-study comparison, and the logical actions to improve reporting in the future.

The remainder of this paper is structured as follows. Section 2 describes the data and methods used. Section 3 documents the problem in four parts, examining how dispersion in the reference group shapes the standardised effect (Section 3.3), how the choice of reference group alters it (Section 3.4), how measurement choice and instrument suitability drive it (Section 3.5), and how secondary influences further complicate interpretation (Section 3.6). Section 4 proposes three practical steps towards improving how interventions are reported, benchmarked and interpreted. Section 5 summarises and concludes.

2. Data and methods

We study evaluations of educational interventions that measure language or literacy outcomes among primary-school children in low- and middle-income countries. This paper does not present a systematic review. Instead, we assemble a large, purpose-built dataset designed to answer this

paper's analytical question, namely how learning effects are measured, standardised, and reported across this literature. The resulting sample represents the peer-reviewed evidence base currently informing policy on literacy and reading, and is intended to capture the variation in measurement and reporting practices most relevant to our argument. Table 1 summarises our sample.

The sample is built primarily from existing evidence reviews. We begin with two large compilations of studies in the education literature, Evans & Yuan (2022, with 234 studies) and the Global Education Evidence Advisory Panel (GEEAP) database (GEEAP, 2023a, with 238 studies).[5] We supplement these with Bertling et al. (2023, with 136 studies), which focuses on learning measurement in LMICs, and Kim et al. (2020, with 65 studies), which reviews literacy interventions in LMICs, together with a non-exhaustive update search for relevant studies published since 2022 (51 studies).

From these sources, we focus on peer-reviewed papers that (i) use an experimental or quasi-experimental design,[6] (ii) measure learning effects for primary-school children, (iii) report at least one reading, language or literacy outcome, and (iv) are conducted in a low- or middle-income country. This yields 171 papers (Annex B details the construction and full paper list).

From each paper's results tables we extract every reported estimate in the units used by the authors. Because papers typically report multiple estimates across specifications, subgroups, and outcome domains, this produces an effect-level dataset of 2,018 measures of a language or literacy outcome.[7] We use these estimates descriptively to document reporting patterns and to illustrate how standardisation choices affect interpretation.

Nineteen of the 171 papers report oral reading fluency outcomes measured in correct words per minute. Because these studies share a common raw metric, the fluency subset lets us compare the same nominal gain across settings. For this exercise, we also draw on Sandefur et al. (2023), who systematically reanalysed USAID-funded Early Grade Reading Assessment studies using the underlying microdata. From their sample we retain the 26 evaluations that used an experimental or quasi-experimental design, report both a standardised and a raw (correct words per minute) effect, and do not already appear in our sample. These programme evaluations were not published in peer-reviewed journals; we use them only in the reading-fluency analysis because they provide paired raw and standardised effects on a common metric, and we exclude them from our description of reporting trends in the peer-reviewed literature.

5 The GEEAP online appendix (GEEAP, 2023b) indicates 235 "Studies used for evidence synthesis" in the PRISMA flow diagram, but the GEEAP study list (GEEAP, 2023c) includes 238 studies which were included in the Full Paper Review and had findings included.

6 We take a relatively relaxed approach to this criterion, including designs such as difference-in-differences, regression discontinuity, and propensity score matching. Our purpose is to examine how outcome effects are measured and reported across the literature, not to estimate a mean effect under a certain identification standard.

7 Including composites where language and another subject are combined into an index.

Finally, we use replication data for 18 studies (14 from the main sample, 4 from the USAID studies) as illustrative case studies. These data allow us to recover summary statistics often omitted in published papers (such as zero-score shares or reference-group standard deviations) and to demonstrate how standardised effects shift when the reference group changes (Section 3.4). We use these files to provide missing descriptive context and to illustrate denominator sensitivity, not to generate new treatment-effect estimates.

TABLE 1. Sample summary

Purpose	Sample Size
Document how papers report gains and how standardisation choices affect interpretation and which measures are used	171 studies: compiled from existing reviews plus an updating search
Compare oral reading fluency gains in raw and standardised terms	45 studies: 19 from the sample of 171 + 26 from Sandefur et al. (2023)
Provide case studies of how reporting and scaling choices can influence effect sizes	18 studies: 14 from the 171 studies + 4 from Sandefur et al. (2023)

3. The challenge: Standardised effects, in isolation, tell us little about meaningful impact

The effect size of any programme is sensitive to many features of how its evaluation is designed, how outcomes are measured and how effects are scaled and reported (Figure 1). These features fall into three broad categories. First, how skills are measured. Second, how tests are scored, standardised and reported. Third, how the intervention design and analysis are specified. Each can materially change the reported effect size and therefore our perception of programme success.

FIGURE 1. Factors that influence reported programme effects

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?

3.1. Known influences on reported effect sizes

The literature documents some of these influences clearly, particularly those relating to intervention design. For example, Evans & Yuan (2022) demonstrate that reported effects are larger for small-scale studies than for large-scale studies, and that quasi-experimental methods generally report smaller effects than randomised trials. Sandefur et al. (2023) offer a potential explanation for this observation, proposing a "tendency for 'better' (and richer) programs to receive more rigorous evaluations." Supplementing this, Kraft (2020) summarises the US evidence as showing that reported correlational relationships are, on average, substantially larger than estimates of causal effects. Other design factors, including how to treat short-run versus multi-year effects and how to compare programmes with different counterfactual conditions, also influence interpretation and are discussed in Section 3.6.

How skills are measured and how tests are scored and standardised are equally consequential for the reported effect size. In high income settings, Cheung & Slavin (2016) show that researcher-designed tests yield roughly double the effects of independent measures. Simpson (2017) highlights how sample and test design choices allow researchers to "legitimately directly manipulate effect size"; and Baguley (2009) argues that standardised effects represent sample variability rather than the meaning of the construct.

In low and middle income settings, a small body of evidence raises similar concerns. Evans & Yuan (2022) show that the form of the test matters substantially, with orally administered tests yielding much larger effects than written tests and multiple choice items larger effects than open-ended items. Singh (2015a) warned that standardised effects are not neutral yardsticks. Bertling et al. (2023) document wide variation in test content for similar grades and subjects, severe floor effects, and limited documentation of test properties. Several authors have called for more comparable or more interpretable outcome metrics (Glewwe & Muralidharan, 2016; Stern & Piper, 2019; Evans & Yuan, 2022). Taken together these contributions establish that measurement and scaling choices matter. What is less well understood is how these choices influence the interpretation of reported effects across a wide range of low and middle income country evaluations.

3.2. Why measurement and test scoring matter for effect-size interpretation

We do not propose replacing standardised effects. But without the raw gain, the reference distribution, and information on what the SD reflects, they can create a false sense of comparability across studies and distort how results are interpreted.

A standardised effect is just a raw gain divided by a chosen standard deviation (see Annex A for a brief description). Change the way skills are measured, who is in the sample, or which group provides the SD, and the reported effect can double or halve even if children's actual learning progress is identical. Standardised effects are comparable when the SD is externally anchored, for example fixed

by biology (as with height) or by calibration to a reference population (as with PISA where 1SD = 100 points by construction).

For most educational interventions however, the SD reflects sample composition and test properties that are internal to the study. The size of a standardised effect isn't just about the underlying relationship between programme and reading skills, it also depends on how much variability there happens to be in the study data. Narrow, homogeneous samples shrink the control-group SD and inflate apparent effects; diverse or noisy samples widen it and dilute them. In other words, the standardisation process upweights progress in less variable groups, a situation that is particularly relevant in low literacy environments (Stern & Piper, 2019). Researcher choices about which reference group supplies the standard deviation (for example, a control group at baseline versus the full sample at endline) also shift reported effects.

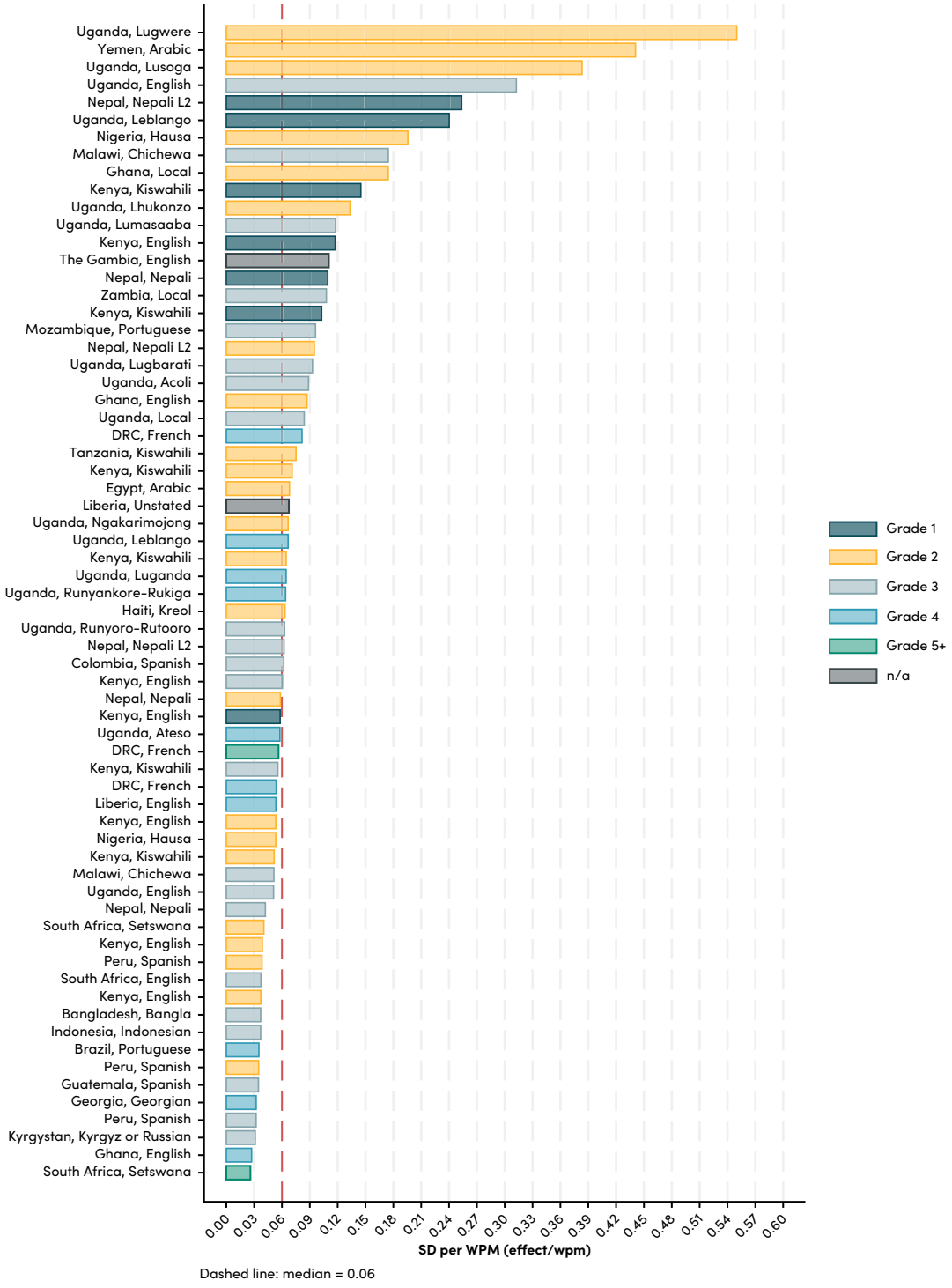
Test properties compound the problem. If a reading test is short, covers a narrow skill domain, or is targeted at the wrong level for the children being tested, the observed SD captures features of that measure, rather than true variation in reading skills. In such cases, SD units can make results comparable in form, but not comparable in meaning.

This is clearest in oral reading fluency data where the raw metric, correct words per minute (CWPM), has a simple and intuitive interpretation. Yet across EGRA studies, one additional correct word per minute corresponds to anywhere between 0.03 and 0.55 standard deviations. (Figure 2 shows this variation across 66 estimates, and Annex Figure C1 provides an alternative interpretation of the same point by flipping the horizontal axis to show the number of words each study would need to reach a 0.3 SD gain.)

The same one-word gain has a six-times larger standardised effect size in a study in Uganda (0.24 SD) than in a study in Kenya (0.03 SD), and is more than twice the size reported in The Gambia (0.11 SD). In a Lugwere-language assessment (Uganda, 0.55 SD) a single word is almost twenty times as valuable as in a Setswana-language assessment (South Africa, 0.03 SD). These are differences between individual studies, not between countries, and this variation does not appear to be primarily a cross-language or cross-grade phenomenon. Figure 2 shows a large range of values within single languages and within single grades. English-language assessments alone span an almost fivefold range across six countries. Holding the grade constant we also see Grade 2 assessments, for example, ranging from 0.04 to 0.55. The effect of one extra correct word per minute cannot differ by factors of six or twenty.[8]

8 Especially if you consider that the Kenya and The Gambia tests were both in English and the Uganda test was in Leblango which has a shallower orthography than English.

FIGURE 2. The number of standard deviations represented by a one word per minute gain varies twenty-fold across studies

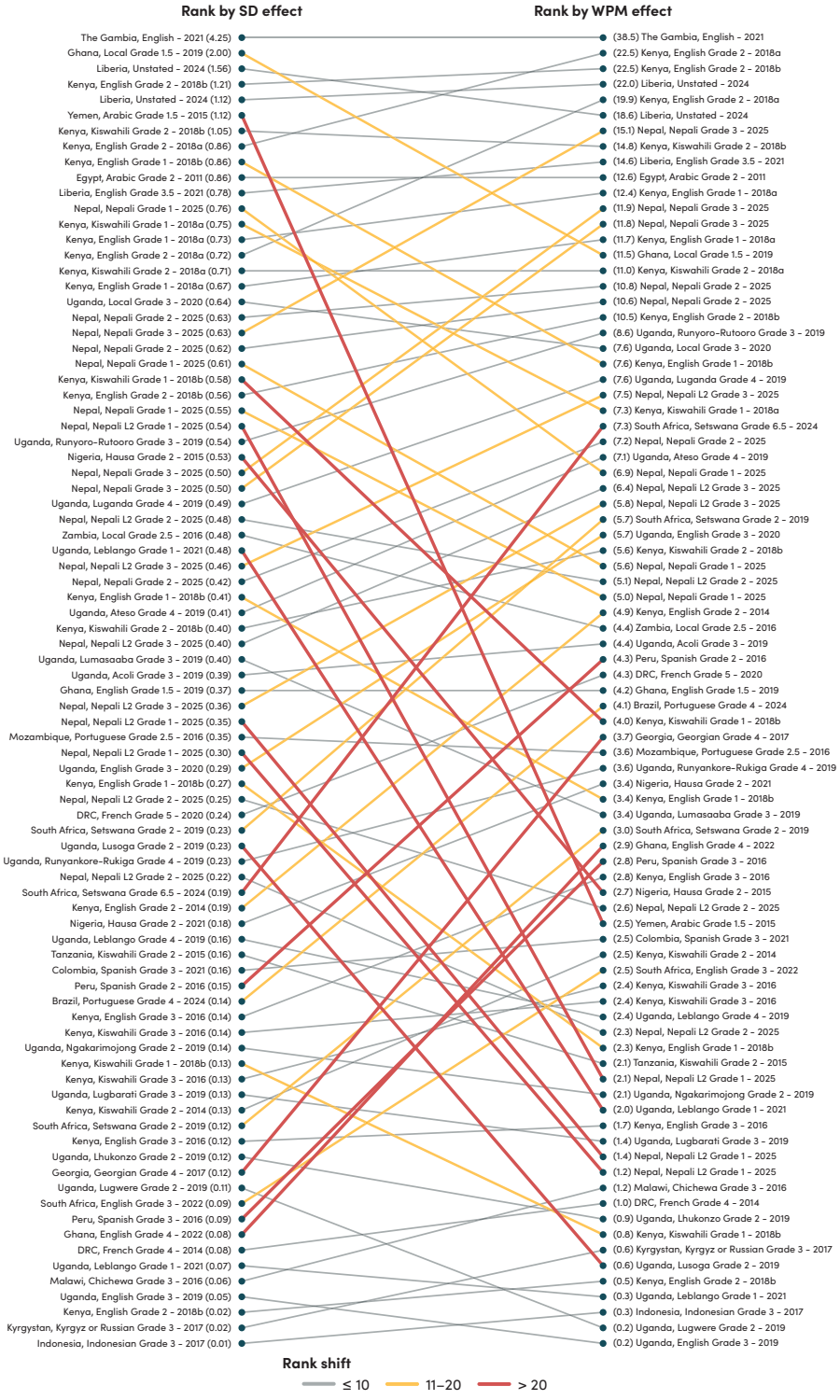


Notes: SD-per-word ratio (SD effect ÷ WPM effect) for 66 study × language × grade cells, each represented by the within-cell median where multiple observations exist. Bars are coloured by grade level; grey indicates grade not available because these two studies worked with household samples. Where a bar represents multiple grades and the average is not an integer we round to the higher grade (8 cases). The dashed line marks the overall median (0.06). The ratio captures how many standard deviations each additional word per minute is "worth" in a given study.

The illusion of variation from a single word per minute arises because the standard deviation used in the standardisation process reflects sample composition, difficulty targeting and language, not a change in what "one more correct word per minute" actually represents in terms of the path to reading comprehension. A simple, interpretable fluency metric is being converted into SD units that obscure what children can actually do differently as a result of an intervention. Heuristic benchmarks for "small", "medium" and "large" effects are of little help here. The range of exchange rates across studies means that a given SD effect (say 0.30) could correspond to a gain of anywhere from half a word to seven or eight words per minute.

The consequences of this variation extend beyond individual studies to the ranking of programmes. Figure 3 compares the rank ordering of 85 observations (wherein both the SD and the WPM effect was greater than zero) when ranked by SD effect versus WPM effect. Despite a strong correlation (Spearman $\rho = 0.82$), individual observations shift substantially. Some of the largest movers are interventions in low-literacy contexts where small absolute WPM gains produce large SD effects. For example, Yemen goes from 5th in SD gains to 58th in WPM gains (top-third to bottom-third) and Nigeria goes from 28th to 56th (top-third to middle-third). Converting raw to SD effects can therefore reorder which programmes appear most effective, with important consequences for policy conclusions drawn from cross-study comparisons.

FIGURE 3. Converting raw to standardised effects reorders programme rankings



Notes: Rank comparison of 85 positive dual-metric observations (SD > 0 and WPM > 0). Each observation is an individual effect. Different treatment arms, grades, and cohorts are not averaged. The left axis ranks by SD effect; the right axis ranks by WPM effect. Lines are coloured by the absolute rank shift: grey (≤10 positions), orange (11–20), red (>20). Nine non-positive observations are excluded.

A further problem is that studies rarely provide the raw effects needed for interpretation. Across the papers in our dataset, the trend has been toward reporting only standardised effects. Table 2 shows how reporting practices have shifted over time. The share of papers presenting any SD-based effect has risen from 37 percent before 2010 to 87 percent since 2020, while the share presenting any raw (unstandardised) effect has fallen from 78 to 43 percent over the same period. Only 23 percent of papers report both metrics. The dominance of standardised reporting is not only a conceptual issue but a practical barrier to understanding what programmes achieved.

TABLE 2. Effect-unit reporting by publication period (171 papers)

Period	Studies	Any SD Effect	Any Raw Effect	Both SD and Raw Effects	Only Other Effects
Before 2010	27	10 (37%)	21 (78%)	4 (15%)	0 (0%)
2010 to 2014	50	30 (60%)	28 (56%)	9 (18%)	1 (2%)
2015 to 2019	47	38 (81%)	19 (40%)	12 (26%)	2 (4%)
2020 to 2025	47	41 (87%)	20 (43%)	15 (32%)	1 (2%)
All	171	119 (70%)	88 (51%)	40 (23%)	4 (2%)

Notes: Classification is based on the units in which treatment effects are presented in the paper's results table or, where the body text includes a discussion of effects in alternative units, in that discussion. Proportion-based effects are grouped with raw (both are unstandardised).

This pattern reinforces the central concern. Without raw effects and sample context, standardisation has the potential to misdirect policy (Simpson, 2017). Policymakers risk allocating resources according to statistical distortions rather than meaningful improvements in children's skills. We examine these challenges in four parts in the remainder of this section.

3.3. How dispersion in the reference group shapes the standardised effect

The standardised effect divides a raw gain by a standard deviation. The spread of scores in the reference group, which supplies that denominator, is therefore as important for the reported effect as the raw gain itself. Where the reference group is homogeneous the SD is narrow and even a modest raw gain translates into a large standardised effect as we illustrate below with two early grade reading programmes. Where scores are more dispersed, the same gain registers as a much smaller one.

The most common source of homogeneity in low and middle income country reading assessments is floor effects, with many children clustered at or near zero. Ceiling effects can produce the same compression when a test is too easy for the target population, though these are less common (Bertling et al., 2023). Oral reading fluency data, where the raw metric is directly interpretable, expose the problems clearly.

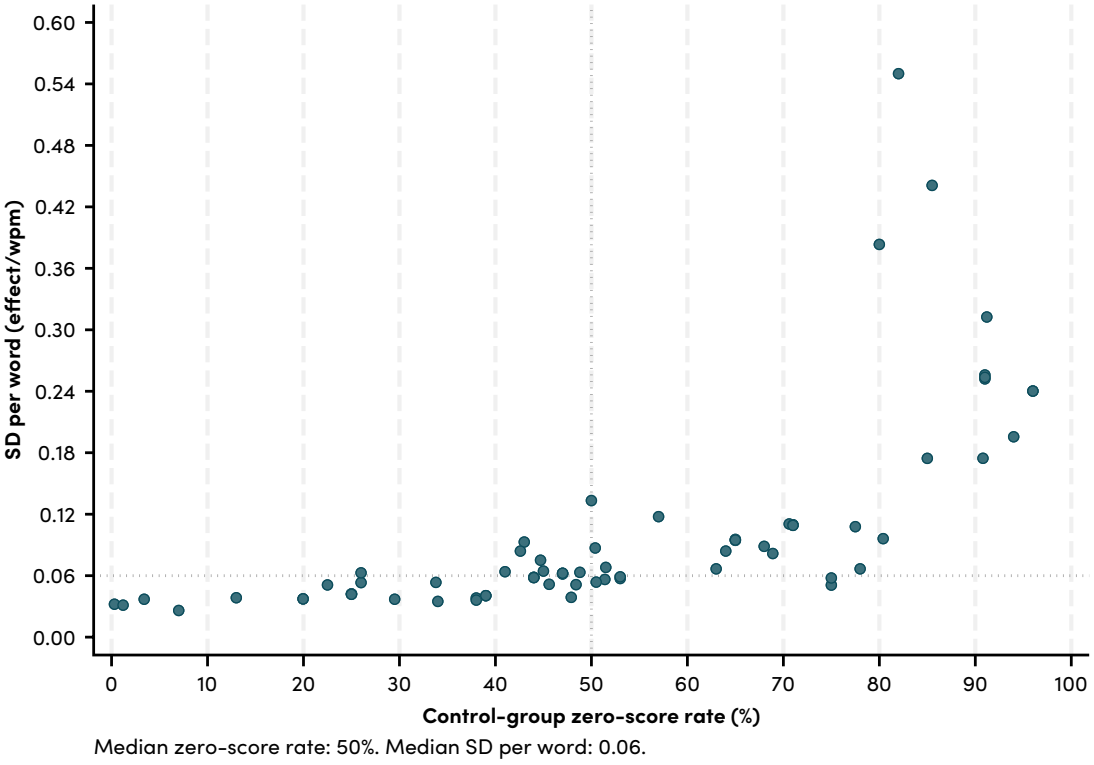
The EGRA oral reading fluency subtask returns scores that are already expressed on a common metric of correct words per minute (CWPM). So, in principle, no further standardisation should be required to compare effects across contexts. Yet many studies convert CWPM gains into standardised effect sizes, aiming to enhance comparability across languages or with studies that have different outcome measures. In practice, this amplifies spurious variation more than it improves comparability.

Among the observations in our combined fluency sample, the pattern is stark. Studies with similar raw gains in words per minute report SD effects that differ by a factor of five or more, driven almost entirely by differences in the control-group SD. Figure 4 shows one feature of the sample, the share of children scoring zero, that underlies this variation.

In the oral reading fluency data, the share of children scoring zero is a strong predictor of reference-group homogeneity and therefore of how large the SD-per-word ratio becomes. Figure 4 plots this relationship for the 71 observations where both a zero-score rate and the SD-per-word ratio are available. As the floor population grows, the control-group SD shrinks and each additional word per minute buys a larger standardised effect ($r = 0.67$ for SD-per-word ratio vs % zeros). The relationship is non-linear. The SD-per-word ratio rises continuously, but the steepening is particularly visible beyond roughly 50 percent zeros, where the few children reading at all provide too little variance to anchor a stable denominator (this threshold is illustrative rather than a formal breakpoint).[9] The shape of the underlying distribution, rather than the intervention itself, drives the apparent size of the effect.

9 This threshold is arbitrary but helps to illustrate the mechanism. The analysis can tolerate a substantial share of zero scores as long as enough children record positive values. When around 40–50 percent of students read at least one word, their scores typically span a broad range, from a few to over a hundred correct words per minute. This provides sufficient spread to stabilise the standard deviation. In that range, shifting a few children from zero into positive territory barely changes the SD. Beyond roughly 70 percent zeros, however, the distribution becomes too truncated. The few positive scores no longer provide enough variation and the SD starts to move sharply with small changes in who clears the zero threshold.

FIGURE 4. More zero scorers in the control group predicts inflated SD-per-word ratios



Note: Each point is an individual effect (a specific outcome, treatment arm, grade or language within a study).
Source: Fluency effect size database (71 observations from 33 studies).

Figure 4 could equally be read as a statement about instrument suitability. An assessment that is too difficult for the target population, or mismatched in language or construct coverage, will generate a large floor effect and compress the reference-group SD. Large standardised effects deserve particular scrutiny in low-literacy settings, where the denominator is most compressed and each raw word buys the most SD units. A case comparison illustrates the consequences.

A comparison between a large-scale literacy programme in Uganda and a comparable programme in Kenya illustrates this pattern well. Both were early-grade literacy programmes combining material inputs with intensive teacher training and support. Both used the same measurement instrument, the EGRA.

In the Ugandan programme, at the end of first grade, mother-tongue letter recognition improved by 1.01 SDs and overall reading by 0.64 SDs. Writing effects ranged from 0.45 to 1.31 SDs across tasks. These effects were described as comparable to some of the largest measured in the literature.

In the Kenyan programme, using the same EGRA tasks, mean effects were still large, but more modest: 0.46 SD across reading subtasks, including 0.73 SD for correct letters per minute and 0.40 SD for oral reading fluency.

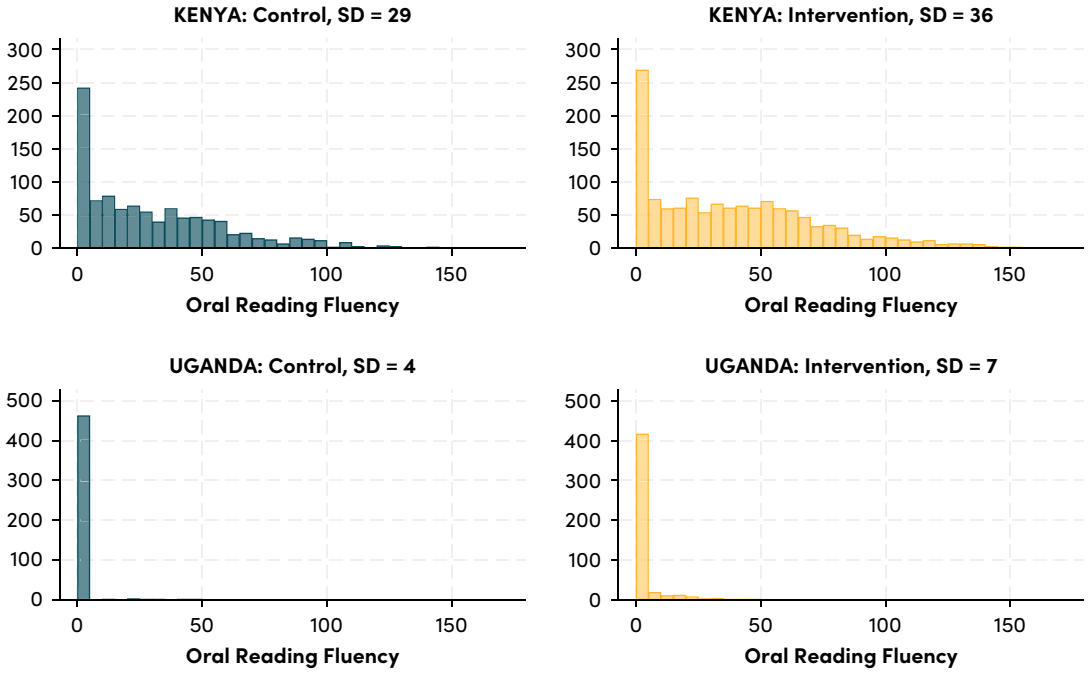
Expressed in raw terms, however, the picture reverses.[10] At endline, children in the Kenyan programme's treatment schools averaged 45 correct words per minute compared with 31 in control schools, a gain of 14 words per minute that moved many learners closer to national benchmarks for their grade (Piper & Mugenda, 2014). They also recognised nearly twice as many letters per minute (47 versus 26 in control schools). In the Ugandan programme, by contrast, gains were just 2 correct words per minute and 10 letters.[11] These much smaller improvements in actual reading ability fall *below* the USAID average for early-grade literacy interventions (Sandefur et al., 2023).

This discrepancy exposes a central challenge in comparing programme effects. The same class of interventions can appear transformative or barely moderate depending on what goes into standardised scores. In the Ugandan case, the large standardised effect reflects that, at the endline, children in control schools had terribly low proficiency on almost all tasks. 46 percent couldn't name a single letter, 97 percent couldn't identify a single familiar word and 96 percent scored zero in oral reading fluency. Only 21 of 473 children recorded any fluency score at all and just 10 could read five or more words per minute. With no meaningful spread of scores, a raw gain of less than 2 words per minute translated into a large standardised effect. In contrast, Figure 5 shows that the Kenyan programme's distribution of reading fluency scores had far fewer zero scorers, providing a more stable basis for interpreting progress.

10 We chose this pair to demonstrate the mechanism because both programmes used the same instrument and a similar intervention model, isolating the role of the reference-group distribution. Other study pairs show different patterns; the point is not that all comparisons are this stark, but that the denominator can dominate the effect.

11 These are calculated from raw scores in replication data, using the same specification as for standardised outcomes.

FIGURE 5. Oral reading fluency rates by treatment group in projects in Kenya and Uganda



Notes: Distribution of oral reading fluency scores (correct words per minute) by treatment status at endline for Kenya (Piper & Mugenda, 2014) and Uganda (Kerwin & Thornton, 2021). In Kenya, both groups show a spread of scores with clear separation between treatment and control. In Uganda, the control distribution is dominated by zero-scorers, compressing the SD and inflating the standardised effect.

Source: Replication data for each study.

The dispersion of scores in the reference group is not a design choice but a feature of the setting, the population, and the instrument. But the researcher does choose which group's dispersion to use as the denominator, and that choice, examined next, adds a further layer of variation to reported effects.

3.4. How choice of reference group alters the effect size

The first issue was the amount of dispersion. The second issue is which group's dispersion we use. Because the standardised effect is constructed by dividing the raw gain by a selected SD, the choice of reference group becomes a powerful lever. Different denominators can produce very different effect sizes even when the learning change is identical. In practice, studies standardise by the average of the baseline control group, the endline control group, a pooled average of treatment and control groups, or subgroup-specific distributions (for example, separate distributions for each grade or gender).

3.4.1. Standardisation conventions and their effects

Small analytical decisions that researchers make when standardising an estimated difference in means can have large consequences for the magnitude of the corresponding effect size (Kraft, 2020). In most cases the choice is a subjective decision, sometimes influenced by the study design and data availability, other times by disciplinary convention.

The most common approach is to standardise the raw effect using the control-group standard deviation, on the logic that variation in the treatment group may be directly influenced by the intervention. When using the control group, researchers may draw the standard deviation from the baseline (where available) or the endline. A related approach is to standardise outcomes separately in each assessment round, producing wave-specific 'z-scores'. For the outcome measure this is equivalent to standardising by the endline control group, and we think of it in that way here.

A variation is to produce group-specific z-scores, by age, grade, location, and so on, before running the analysis. This removes group-driven differences in means and variances, yielding a cleaner within-group measure of learning gains and avoiding composition bias. Yet it also changes the interpretation of the standardised effect size, moving it away from "how much did the programme shift the full sample compared to the control overall?" to "how much did it shift children at the same developmental stage? / in the same location?".[12]

The other common approach is to use Cohen's *d*, the pooled standard deviation across both control and treatment groups at endline. This approach treats the variation in each group as equally relevant to interpretation.

Across the 119 papers that report standardised results, we see a clear tendency to favour the control-group standard deviation at endline (Table 3), but all approaches are used. Standardisation is a paper-level choice with every paper in our dataset using a single denominator group across all its SD effects (i.e., we do not see switching between using a pooled group for some effects and the control group for others).

TABLE 3. Standardisation denominator choices (119 papers reporting SD effects)

Standardised By	At Endline	At Baseline	Unspecified	Total
Control/comparison group	57	5	5	67
Pooled/full sample (Cohen's *d*)	18	8	6	32
Not specified or ambiguous	0	0	20	20
Total	75	13	31	119

Note: Eight papers standardise within sub-populations such as age groups, grade cohorts, language groups, or examination years rather than against a single pooled or control-group distribution.

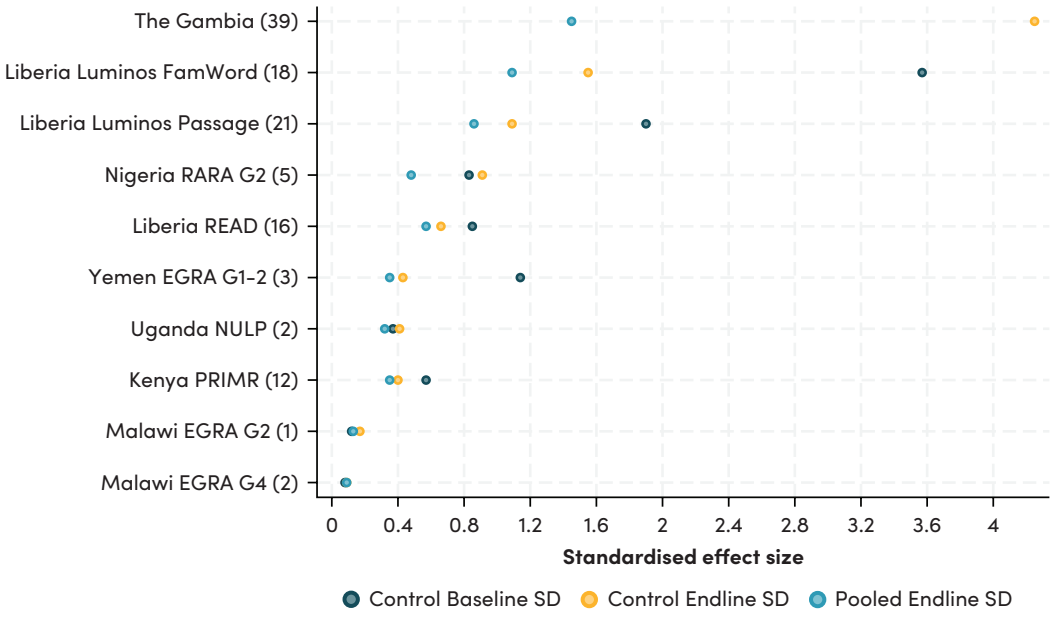
This choice is not trivial. The sample used to calculate the SD determines the denominator of the effect size. In plain terms, a 0.4 SD effect is the same as saying that the intervention group mean increased by 0.4 units *of the relevant distribution of scores*, relative to the control group mean. Changing the reference distribution of scores alters the standardised effect size even when the underlying raw improvement is identical.

12 More technically, the estimate becomes an average within-group SD gain rather than an SD gain relative to the overall control distribution.

The magnitude of difference between different approaches is not consistent across studies (Figure 6). In some cases, the choice of reference group barely alters the reported effect (see Malawi, or Uganda), while in others it transforms the interpretation (see Yemen or The Gambia). Despite the horizontal scale being driven by the largest effect in Figure 6, for most of these studies, the differences are very large and will influence how readers judge intervention performance: from about 0.4 to over 1.0 in Yemen, from 0.5 to 0.9 in Nigeria, from less than 0.4 to almost 0.6 in Kenya.

This variation reflects three interacting features, namely the raw programme effect, the relative variance of treatment or control groups at endline and what it was that the control group was exposed to while the treatment group received the intervention. When the control group remains largely inactive or at the floor of performance, standardisation choices are particularly influential, but when both groups are, say, in school and making some progress, the standardisation can be less important.

FIGURE 6. Reference group choices can have large impacts on standardised effects



Notes: This figure shows standardised effect sizes that result from analyst choices about which distribution to use as the reference. All estimates are based on simple mean differences from replication data, using design weights where available. In The Gambia, there was no baseline assessment so only two values are shown. For each study, the raw difference in correct words per minute is shown in parentheses.

We don't think that there is necessarily any single "correct" reference distribution, but the choice should be explicit. It is worth distinguishing the two purposes a denominator serves. For within-study inference, the endline control group is the natural choice because it represents

the counterfactual. Where a within-study default is needed, we recommend it on that basis, while recognising that other variants can be defensible depending on design and research intent.[13]

For cross-study synthesis, however, even a uniformly applied convention cannot resolve the comparability problem. Two studies that both divide by the endline control-group SD will still produce denominators shaped by different sample and instrument features, differences large enough, as we have shown, to reverse apparent programme rankings. Cross-study comparability therefore depends less on which denominator is adopted and more on whether the raw effects, score distributions, and instrument characteristics needed to model these differences are reported alongside the standardised figure.At the very least, authors should clearly state the reference group (e.g., "0.2 SD of the distribution of scores in the endline control group"), show the underlying distributions, and test sensitivity to alternative choices. Once the raw score distributions are visible, readers can judge what a given SD effect means in real terms, regardless of which variant is used. The priority is not enforcing a single standard but ensuring that SD-based effects do not give the impression of being comparable when they rest on different denominators.

3.4.2. Subgroup analysis as a reference-group choice

Researchers sometimes report results separately by grade, language, school type, or location. This disaggregation is valuable for policy since it can reveal who benefits the most from a programme, but it also constitutes a reference-group choice. When each subgroup is standardised against its own control distribution rather than the overall control, the denominator changes across groups. As subgroups become more homogeneous their SDs narrow and the same raw gain translates into a larger standardised effect (Baguley, 2009).[14]

Table 4 illustrates this with data from the Primary Math and Reading (PRIMR) Initiative. The same five-item reading-comprehension outcome yields quite different standardised effects depending on which slice of the sample provides the denominator.Grade 1 and Grade 2 formal students show similar raw gains, yet their standardised effects differ by a factor of 1.7 (0.58 vs 0.35). In Grade 2, one additional correct response counts as 1.09 SD in formal schools but only 0.76 SD in nonformal schools, a 45% premium simply because one group has a narrower distribution. These shifts come entirely from a choice to standardise within subgroups rather than against the full control.

13 Baseline control-group SDs may be preferred where attrition or contamination is a concern at endline. Pooled (treatment + control) SDs may be preferred for comparability with Cohen's *d* conventions or where the treatment is expected to change the variance. The key requirement is not that everyone uses the same denominator, but that the choice is stated explicitly and sensitivity to alternatives is tested.

14 A more general source of range restriction comes from studies that operate in sub-regions or prioritise certain sub-groups in a country. If a subset of rural schools has been included in the study then it is likely that the range of performance in this group is considerably narrower than at the country level. Range restriction is also influenced by data processing choices, for example in removing suspected outliers from datasets, or winsorizing scores. This mechanically reduces the range of observed scores in the analysis, reducing the standard deviation.

TABLE 4. Effect sizes for a series of subgroups from the same overall sample

Group	Grade 1 Formal	Grade 1 Nonformal	Grade 2 Formal	Grade 2 Nonformal
Raw effect on reading comp.	0.38	−0.05	0.32	0.21
Control standard deviation	0.66	1.24	0.92	1.28
Standardised effect size	0.58 SD	−0.04 SD	0.35 SD	0.16 SD
Group	**Grade 1 Overall**		**Grade 2 Overall**	
Raw effect on reading comp.	0.19		0.26	
Control standard deviation	1.03		1.12	
Standardised effect size	0.18 SD		0.23 SD	
Group	**Grade 1 & Grade 2 Overall**			
Raw effect on reading comp.	0.23			
Control standard deviation	1.09			
Standardised effect size	0.21 SD			

Notes: The "Raw effect on reading comp." is the difference between intervention and control groups' average score on five reading comprehension questions. Weights are used to compute means and standard deviations. We standardise using the control standard deviation for the relevant group (i.e., the control in G1 formal schools for the "G1 Formal" group; or the full G1 control sample for the "G1 Overall" group).

Source: Replication data for PRIMR midline.

Claims such as "the programme worked especially well for rural girls" can therefore describe data structure more than educational reality. The safeguard is to standardise once against the full reference distribution, so that all effects are expressed relative to the same reference SD. Subgroup analyses should then report both raw and standardised effects alongside sample sizes, allowing readers to judge whether differences reflect genuine heterogeneity or result from the mathematics of smaller denominators.[15]

3.5. Measurement choice and instrument suitability

So far, we have focused on how variation within and between groups affects the scale against which progress is judged. A third challenge lies in how that progress is measured in the first place. Even with identical underlying gains in skills, two studies can report very different raw and standardised effects simply because their instruments capture different constructs, target different difficulty levels, or are better (or worse) aligned to the children being tested.

In this section, we describe how instrument fit can drive the standardised effect size. By instrument fit we mean the match between what a test measures, the ability level of the children being assessed, and the progress it can and cannot detect. Two questions are central:

- What construct or skill domain does the instrument cover?
- How well does its difficulty match the ability level of the learners being assessed (especially, how severe are the floor and ceiling effects)?

15 Piper et al. (2014) does this excellently for subgroups in PRIMR, with Table 4 showing simple comparison of means by subgroup, along with effect sizes and standard deviations used to compute those.

Understanding why measurement suitability matters for reading requires briefly considering what reading development looks like. Reading for meaning rests on a developmental sequence of subskills. Children must first master letter-sound associations, then use those sounds to decode new words, and eventually read sentences fast enough that working memory can assemble meaning. The fluency threshold at which comprehension becomes possible varies across languages, depending on orthographic depth and morphological structure (Section 3.6.4 discusses these language-specific differences in detail). What matters here is that any single instrument, whether it measures letter knowledge, decoding, fluency, or comprehension, captures only part of this sequence.

Foundational skills such as letter identification, letter-sound knowledge, blending, and decoding are strong predictors of later comprehension and represent meaningful progress in their own right. Crawford et al. (2024) show that children in LMICs make real gains in these early skills between Years 1 and 3, even when they remain far from fluency benchmarks. A child who learns 15 new letter-sounds or begins decoding simple words has crossed a critical developmental threshold, yet a fluency-based measure would record zero progress. This mismatch between what children have learned and what the instrument can detect is a problem of instrument fit and is the central concern of this subsection.

3.5.1. A fragmented measurement landscape

The EGRA is the only reading assessment that lets us directly compare nominal gains (raw words per minute) with standardised effects across many studies. Yet it appears in just 11 percent of papers in our dataset (Table 5). With the dismantling of USAID and its funding of early grade reading studies (Walls, 2025), the sector will likely confront an even more fragmented measurement landscape.

TABLE 5. Assessment types used across 171 papers, by publication period

Period	N	Gov. Exam	EGRA	Other Vendor	Researcher	Multiple
Before 2010	27	9 (33%)	0 (0%)	2 (7%)	16 (59%)	0 (0%)
2010–2014	50	13 (26%)	0 (0%)	3 (6%)	33 (66%)	1 (2%)
2015–2019	47	6 (13%)	7 (15%)	3 (6%)	28 (60%)	3 (6%)
2020–2025	47	6 (13%)	11 (23%)	2 (4%)	26 (55%)	2 (4%)
All	171	34 (20%)	18 (11%)	10 (6%)	103 (60%)	6 (4%)

Nearly two-thirds of papers use researcher-designed instruments; the rest rely on tools from government exams, EGRA, or other named tests such as the ASER and PPVT measures. These instruments measure different constructs, vary in difficulty, and present children with tasks that are rarely comparable. This fragmentation introduces a further complication. Nominal changes on different tests cannot be easily compared, but their standardised effect sizes routinely are.

Kerwin & Thornton (2021), whose Uganda data we examined in Section 3.3 for reference-group dispersion, provide further evidence on this point. They show that within a single randomised trial

the apparent effectiveness of a literacy programme can shift dramatically depending on which outcome measure is chosen. As they note, "there are many possible measures of learning: a wide range of tests, measuring a variety of skills and implemented in different languages." Small changes in test construction, such as using letter-recognition versus sentence-writing tasks, can alter the measured effect by more than half a standard deviation. Their findings highlight how both input choices and measurement design jointly determine observed programme impact, even when implementation and context are held constant.

Within the Northern Uganda Literacy Project, for example, effect sizes ranged from 0.14 SD (story-writing, presentation) to 1.31 SD (given name writing) across EGRA reading subtasks and writing subtasks administered in the same trial. Even summarising across subtasks, the simple mean of subtask effects (0.53 SD) differed from the PCA-weighted composite (0.64 SD). The choice of which subtask to highlight can therefore shift the headline result by more than half a standard deviation.

Across three studies that use composite language scores, we see the same pattern. These three studies were selected because each constructs a composite language score from reading and writing subtasks, allowing a like-for-like comparison of how similar nominal progress translates into different standardised effects. In Nigeria, small nominal gains (1.5 additional letters sounded, 0.3 more words per minute, near-zero improvement in comprehension) translate to 0.46 SD,[16] enough for the authors to describe the intervention as "highly effective and cost-effective." In Kenya, slightly larger absolute gains (3.5 more letters, 0.6 more decoded words) yield only 0.13 SD.[17] In Colombia, still larger gains (4.4 letters, 2.5 more words per minute) produce 0.27 SD.[18] The range of 0.13 to 0.46 SD for qualitatively similar underlying progress illustrates how much of the apparent variation comes from the test, the sample, and the scaling, rather than from differences in what children learned.

For researcher-designed item-based assessments, the variation is equally stark. A single additional correct response corresponds to anywhere from 0.08 to 0.80 SD across studies in Burkina Faso,[19]

16 In the baseline sample, only 116 of 9,285 children (1.2%) proceeded beyond the letter identification subtask, which required them to give at least one correct answer from the first 10 letters. The authors construct a composite measure of all literacy skills (0.46 SD effect) and calculate that this represents 0.75 LAYS / $100 ranking in the top quarter of 72 comparator interventions.

17 ETP replication data. Comparing raw scores by tracking status and regression of normalised literacy score (control normed) on tracking.

18 All subtask scores are added together, so an additional letter is worth the same as an additional word, same as an additional comprehension question. The total difference was 7.7 'correct responses'. Good example of the aggregation issue.

19 Children were asked to identify 2 letters, decode 4 words, and fill in 2 blanks. The test was the same irrespective of child age. 97% scored zero at age 6, 70% scored zero at age 7, 58% scored zero at age 8.

Bangladesh,[20] Afghanistan, and Indonesia,[21] reflecting differences in samples, constructs, item difficulty, and how narrowly the test is targeted.

Differences across studies are compounded by differences within studies. In both Burkina Faso and Afghanistan, the SD value of a single item depends on the age of the children or the timing of the test. This follows directly from the standardisation process. In the first case, tests are standardised by age, while in the second case, the same assessments are re-standardised at two different points in time.[22]

3.5.2. How instrument properties mechanically alter the standardised effect

This diversity of instruments would be less concerning if different measures yielded similar effect sizes for the same underlying learning gain. But they do not. To illustrate, we consider what happens when the same improvement in children's latent reading ability is captured by two different measurement tools in two different literacy environments.

The first tool is a continuous measure of oral reading fluency, returning correct words per minute (CWPM). The second is an ASER-style categorical ladder classifying each child into one of four levels (cannot identify letters, can identify letters, can read words, or can read a short paragraph). We compare each tool in a higher literacy environment (average fluency ~40 CWPM at baseline) and a lower literacy environment (average fluency ~2 CWPM at baseline).[23]

Although the true learning gain is identical across all four scenarios, the observed standardised effect sizes range from 0.35 to 1.47 SD (Table 6).

TABLE 6. The same real improvement (1 SD) in latent reading ability produces very different observed effect sizes depending on instrument-environment fit

Measure	Latent Ability	Control Mean	Intervention Mean	Mean Difference	SD of Control	Std. Effect
CWPM	higher	40.9	60.6	19.7	20.2	**0.98**
CWPM	lower	1.5	8.5	6.9	4.7	**1.47**
Ladder	higher	3.82	3.98	0.16	0.46	**0.35**
Ladder	lower	1.67	2.51	0.84	0.76	**1.11**

20 Note that it was not 1 more item, in fact it was the same score over time in the intervention group, just that the control had fallen back by 1 item. Three of the six items were translations of Bangla to English words.

21 Pradhan et al. 2014. Based on 0.17 SD in Table 5, "Linkage" estimate, and post-SD in control (Col (1)). The 0.16 SD value is from replication data, comparing raw scores with z-scores.

22 A further issue is treatment-inherent measurement. Researcher-designed tests are more likely to mirror the content or formats used in the intervention, increasing the chances that children in the treatment group perform well because the test aligns with what they practised, not because of broader gains.

23 The illustration assumes that latent reading ability maps linearly to observed CWPM, with baseline means of approximately 40 CWPM (higher literacy environment) and 2 CWPM (lower literacy environment), plus normally distributed measurement noise. The ASER ladder scores are derived from the same latent ability using threshold cut-points.

In the higher literacy environment, reading fluency tracks the true effect closely (0.98 SD), because the tool is well matched to children's abilities, and scores are spread across a wide range with minimal floor effects. But the categorical ladder, designed for lower-literacy contexts, suffers a severe ceiling effect (most children are already at the top level) and captures only 0.35 SD.

In the lower literacy environment, the pattern reverses. Reading fluency produces a raw gain of only 6.9 CWPM (versus 19.7 in the higher-literacy setting), because most children cluster at zero and genuine progress in letter identification and decoding goes undetected beneath the floor. Yet when this small raw gain is standardised against the very narrow control-group SD, the effect inflates to 1.47 SD, well above the true 1 SD. The ladder is more sensitive here, capturing movement across proficiency levels, but still returns an inflated 1.11 SD once standardised.

3.5.3. Empirical evidence: progress hidden beneath the floor

An in-school camps intervention in Uttar Pradesh illustrates this pattern with real data (Banerjee et al., 2017). Of 1,725 children who could not identify letters at baseline, the endline showed:

284 children	(16% of them)	remained unable to identify letters
810 children	(47% of them)	had progressed to letter identification
311 children	(18% of them)	had progressed to word identification
310 children	(18% of them)	had reached a level that they could read a short paragraph or more, i.e. some measurable fluency

Sole reliance on a reading fluency measure would have overlooked progress for approximately 1,500 children (a third of the total intervention sample) who had demonstrably improved their reading ability but had not yet reached the threshold of measurable fluency. A further 1,794 children who could identify letters but nothing more at baseline showed similar patterns. Forty-two percent remained below the level at which a reading fluency measure would detect any change.

3.5.4. Implications for instrument choice and reporting

These findings reinforce the central point. Effect sizes from different studies are not inherently comparable unless the measurement context and instrument design are taken into account. Oral reading fluency works well where children are already reading but misrepresents progress in low-literacy environments where many children sit at the floor. Categorical tools capture early gains but saturate quickly in more literate settings. Standardisation amplifies the problem in both directions, inflating effects where distributions are compressed, and attenuating them where ceiling effects dominate.

Instruments that concentrate most of the sample at the floor or ceiling will generate misleading effect sizes and should not be treated as interchangeable with well-targeted measures. Authors should show the raw score distributions so that readers can judge what a given SD effect means

in real terms and how suitable the measurement tool was. Where the scale has no natural interpretation (for example, because scores are produced using IRT), researchers should explain what differences on that scale represent for children's skills. Following the guidance set out in Bertling et al. (2023) (piloting assessments in context, ensuring adequate variation across individuals, documenting psychometric properties) would avoid most of the problems illustrated here and make subsequent effect-size comparisons far more trustworthy.

These three channels, reference-group dispersion, denominator choice, and instrument suitability, account for the largest mechanical distortions in standardised effects. But several other features of how evaluations are designed, scored, and reported can also shift the reported effect, and we turn to these next.

3.6. Secondary influences on interpretation

Several other features of study design, test scoring and reporting also influence the size and meaning of standardised effects. Some, particularly scoring conventions and psychometric quality, share territory with the measurement-suitability requirement in 3.5 but operate through different channels. None is uniformly important across all studies, but any one of them can matter in specific cases. We discuss six, organised around how tests are scored and reported, properties of the assessment and features of the design.

3.6.1. How items and subtasks are scored and combined

Aggregation is not a neutral step. Two analysts using the same items can produce different effect sizes purely through their scoring and combination choices.

Instruments such as the ASER ladder do not permit aggregation, producing simple but coarse outcomes (a child is placed into one of four categories). EGRA and its variants allow aggregation but do not specify a standard method. Across our database EGRA subtasks have been combined in several ways.

- Taking a simple average of subtask scores, sometimes after converting to percent correct.
- Summing all items across subtasks into an overall proportion correct.
- Computing weighted averages using researcher-defined, Principal-Component-Analysis, of factor-analytic weights.
- Converting subtasks to binary values and applying item response theory.
- Aggregating some subtasks without specifying how, or excluding certain groups (such as zero scorers).

The central issue is whether the composite reflects a coherent construct. If every item gets one point, identifying the letter *m* counts the same as answering a comprehension question. If each subtask is converted to percent correct and averaged, a 20-item letter-sound task carries the same weight as a

5-item comprehension task. Comparability across studies depends not only on what was measured but on how tasks were aggregated.

Item response theory does not resolve this. IRT works well for independent accuracy items (Kolen & Brennan, 2014) but cannot accommodate rate measures (such as words per minute) or the inter-item dependencies in timed tasks, where success on later items is conditional on earlier ones.[24] For this type of task-based assessment there is no principled method for combining all components into a single construct-valid scale. Best practice is to report effects separately for key subtasks, including in raw units and to make explicit the construction of any composite and its relationship to the intended reading construct.

For researcher-designed instruments the same issues arise. Items are typically summed into a total, but the weighting approach changes the score distribution. In one case, a normalised raw score produced an effect of 0.32 SD while the same data expressed as an IRT-scaled score produced 0.66 SD. An identical learning gain doubled purely by the method of combination (Singh, 2015b).

3.6.2. Which subtasks are reported and emphasised

When an EGRA-type instrument is used, researchers choose not only how to combine subtasks but which to report and emphasise. Since each subtask taps a different component of reading, selecting one over another changes the construct being evaluated. We noted in Section 3.5 how subtask choice can shift the headline effect by more than half a standard deviation within a single trial.

A related issue is interpretability. Many scales have no natural criterion reference. A point on an IRT scale, or on a 30-point reading test, reflects a child's position in a distribution, not mastery of a specific skill. Where scales lack direct meaning, researchers should provide the interpretation, for example by showing what children at different points on the scale can do. Other approaches include using percentage-correct summaries for items linked to a single competency or item-by-item examples that illustrate the practical difference implied by the standardised effect. The virtue of this is that researchers engage directly with the underlying changes in skills and what these mean for children. But it is very rarely done.

3.6.3. Psychometric quality

Whether an assessment is internally coherent, reliable across administrations, and valid for the intended construct affects the credibility of effect estimates. If items within a domain show little common variation across children, summary scores will be unstable. Where scores fluctuate across

24 In principle, each subtask could be treated as a "testlet", with the testlet score defined as the proportion correct, or e.g., fluency rate. But once subtasks are treated this way, calibration requires more complex IRT models and the assessment collapses to only a handful of "items", making estimation unstable, particularly in small samples. More importantly, this approach does not address the problems of timed tasks which violate local independence assumptions. We have not seen any study in our database attempt such testlet-level calibration.

assessors or test occasions, treatment effects are attenuated, and comparisons across groups or rounds become difficult to interpret. These issues are particularly relevant for early-grade assessments relying on oral administration.

Construct validity raises a particular concern. In most EGRA reading comprehension administrations, children can only reach the highest scores if they read the entire passage within the time limit, tying the comprehension score to decoding speed (Dowd & Bartlett, 2019). Reported comprehension effects may therefore partly reflect improvements in fluency. This does not make the assessment unusable, but it influences how effects should be interpreted. More generally, Bertling et al. (2023) find that only 4 percent of researcher-designed tests in their review report reliability estimates. Where psychometric properties are weak or undocumented, effect estimates risk mischaracterising programme impacts regardless of how scores are standardised.

3.6.4. Assessment language

Languages differ in their writing systems, in how consistently letters map to sounds (orthographic depth), and in how words are built from smaller meaningful units (morphology). These differences directly affect how many words a child can read per minute without necessarily indicating higher or lower comprehension (Spaull et al., 2020; Ardington et al., 2021). A words per minute benchmark therefore represents different levels of proficiency across languages (Dowd & Bartlett, 2019). In languages with predictable letter-sound correspondences (transparent orthographies), such as Setswana or Kiswahili, comprehension begins at roughly 35 to 45 CWPM (Jukes et al., 2020). In less predictable languages like English, it typically requires 60 to 70 CWPM (Abadzi, 2012). Word structure also matters. In languages that join meaningful units into longer words (conjunctive orthographies) such as isiZulu a reader may appear slower in words per minute even though each word can carry more meaning (Spaull et al., 2020). For instance, in isiZulu, the sentence “Abantwana bayafunda izincwadi” (“The children are reading books”) combines several grammatical units into one word, while in Sepedi the equivalent “Bana ba a bala dibuka” separates them.

These differences are compounded when children are assessed in a colonial language (English, French, Portuguese), often taught as a second or third language. In transparent orthographies, learners can decode words they do not understand, producing what appears to be fluency without comprehension. This “word calling” inflates apparent fluency relative to understanding (Dowd & Bartlett, 2019; Abadzi & Centanni, 2020). In opaque orthographies like English, the opposite applies. Accurate pronunciation often requires comprehension, since words such as “bow” and “lead” can only be read correctly when the reader understands the context. Fluency and comprehension are therefore more tightly linked in opaque languages and more easily separated in transparent ones. By contrast, small fluency gains in mother-tongue programmes can yield large comprehension improvements, particularly when the language of the assessment aligns with the learner’s oral vocabulary.

3.6.5. What the comparison group experiences

The estimated effect of any programme is a contrast with the comparison group. Variation in what control children experience therefore alters both the magnitude and interpretation of the effect. Consider two cases:

1. A programme provides teachers with technology in existing schools. The sample is drawn from enrolled children. Children in control schools receive regular instruction. The effect measures the additional gain from the programme above regular schooling.
2. A programme establishes new community-based schools in areas with low enrolment. The sample is drawn from households before schools open. Children are randomly assigned. Those not selected may enrol elsewhere or receive no schooling at all. The effect measures the gain from schooling itself, not just from the specific programme.

An effect of 0.3 SD means something different in each setting. The first comparison is against active instruction with a school-attending sample. The second is against a potentially unschooled counterfactual with a population-based sample. When interpreting effects, and especially when judging external validity or scalability, both the counterfactual conditions and the sampling frame should be clear.

3.6.6. Exposure duration

A 0.5 SD (or 6 CWPM) gain achieved over four years signals a different pace of learning from the same gain achieved over one term. Yet it is often hard to tell how long the measured cohort was actually exposed, as opposed to how long the project ran or the time between baseline and endline. Some multi-year programmes follow the same children, so exposure matches programme duration. Others support teachers across several years but test different cohorts who may have received only one year of treatment.

Table 7 links reading comprehension gains to exposure information for five programmes.[25] Those with longer timelines reach the highest total gains but may deliver slower annual progress. Guinea-Bissau records the largest four-year gain (3.54 points, equivalent to 0.89 per year), while the one-year Liberia intervention delivers a smaller total gain but the fastest annual pace (1.20 points per year).

Learning gains are unlikely to be linear and short intensive programmes should not be extrapolated over longer periods. For interpretation, what matters is not just how much children gained but how fast, and that requires clear reporting of exposure duration.

25 We use these programs because they provide a range of exposure durations and because they were effective enough to have made relatively large changes to children's reading comprehension skills, something that is much less common in short term interventions.

TABLE 7. Reading comprehension raw effects for five programmes, by total gain and year of exposure

Intervention Country	Comparison Group Experience	Intervention Duration (Exposure Years)	Reading Comprehension Score at Endline (Correct Responses/5)		Reading Comprehension Points Gained in Intervention Over Control	
			Intervention Mean	Control Mean	Per Intervention	Per Year of Exposure
Guinea Bissau	In control communities	4	3.60	0.06	3.54	0.89
The Gambia	In control communities	3	2.40	0.15	2.25	0.75
Liberia	In control communities	1	1.60	0.40	1.20	1.20
Kenya	In control schools	1.5	1.74	0.97	0.77	0.52
Liberia	In control schools	2	1.60	0.90	0.70	0.35

Notes: Reading comprehension is measured as the number of correct responses to five questions after children attempt a short passage, using EGRA or closely related instruments. "Per year of exposure" divides the total gain by the number of exposure years.

Source: Paper or paper replication data.

4. Three steps toward a more meaningful interpretation of programme effects

This section turns to the practical question of how programme effects should be reported so that readers can tell what children actually gained? As the introduction sets out, we do not propose abandoning standardisation but adding to it with information researchers already collect.

We suggest three steps. First, show the raw gains and score distributions behind every standardised effect, so readers can see what those numbers represent (Step 1). Second, use benchmarks to put gains in perspective, showing how many children crossed a level that matters for their reading (Step 2). Third, use measures that are suited to the skills being tested, so that reported changes reflect real progress rather than features of the test (Step 3).

4.1. Step 1. A request for greater transparency

The first path calls for greater transparency in how programme impacts are reported. Specifically, it asks researchers to report raw programme effects and simple, clear metrics alongside standardised ones. The underlying question we should be able to answer from any study is straightforward:

> *"What is it that children can now do, thanks to the programme?"*[26]

26 Kraft (2020) sets out ten questions to ask when interpreting evaluation findings in US education literature. Several of these would also be helpful, but we focus on the elements that are particularly influential in low- and middle-income countries, which can differ from those in the US or other relatively high-literacy environments.

As we have shown, without the raw gain and the reference distribution, these effects are hard to read as evidence of real learning. Researchers should report two things alongside any standardised effect. First, the unstandardised ("raw") effects, in clear skill terms. Second, the reference sample distribution used to calculate the standard deviation. Reporting what programmes helped children to do in raw terms can then be built into judgments of how well a programme worked.

This path can be implemented immediately. It requires no new data and can be applied retrospectively, as illustrated in the examples in this paper. To support this, researchers should document four things:

1. What distribution are we standardising to when constructing the standardised effect size?
    a. The control group or the full, pooled, sample?
    b. At what point in the intervention, baseline, endline, both?
    c. Are there adjustments by grade, by gender, or by other grouping?
2. What does that raw distribution that we are standardising on look like?
    a. What are the mean and standard deviation for the distribution?
    b. What does the distribution look like?
3. Is the standardised effect size robust to the choice of reference group?
    a. If effects are disaggregated, what would the aggregate/overall effect be?
    b. If standardised effects are derived from a baseline (or endline) group, what would change if we'd made a different choice?
4. What does each point of progress along that distribution mean in real terms?
    a. Where there is a direct interpretation, such as in CWPM, provide that.
    b. Where the scale has no direct interpretation (such as in the case of a percentage correct or IRT score), provide information on what additional things children can do thanks to the programme's impact.

These are minimal requests needed for fair reading of findings, and they add to other information on intervention design (identification strategy, sample size and structure, timeline) that is often well reported. They also push researchers to engage with the meaning of programme effects and guard against publication pressures that can lead to overstating sometimes trivial effects. If we saw these pieces of information alongside the standardised effect size we would quickly have a much better picture of what programmes are doing for children.

Different readers will draw on different pieces of this information. For policymakers and donors, the pairing of a raw effect alongside any standardised one is sufficient to judge whether a programme delivered something meaningful. The more detailed elements can feed into improved evidence aggregation. In a systematic review or meta-analysis, characteristics of this kind can be used as covariates in a meta-regression to test, for example, whether reported effects shrink once differences in test type, share of zero scorers, or choice of denominator are taken into account (Deeks et al., 2024).

4.2. Step 2. Use benchmarks to put gains in perspective

Where we use a measure of reading fluency, we may prefer to focus on unstandardised (raw) effects which readers can understand directly. These can be linked to benchmarks which help put results in perspective and set realistic expectations (Stern & Piper, 2019). For instance, reporting what share of children crossed a key comprehension threshold offers a clearer picture than citing mean fluency or standard deviations. Even when comprehension gains remain modest, benchmarks show where learners stand relative to useful skill levels. They can also show whether progress is concentrated among beginning readers or those near fluency, guiding teaching decisions.

Reporting clear outcomes, rather than standardised effect sizes, can be particularly helpful in communicating with policymakers. A minister of education or finance does not (usually) need a discussion of the choice of denominator or the shape of the score distribution to decide whether a programme is worth looking into. They need to know whether more children are reaching relevant skill levels. Benchmarks do the job of translating a reported effect into more familiar units.

This principle of benchmarking can apply to any assessment measure. It is easiest to put into practice with tools like EGRA, because its tasks produce units (words per minute, letters per minute) that already have a clear meaning. As we have shown, however, the EGRA tool is far from dominant in our dataset. A complementary approach is to anchor results to descriptions of specific reading skills or competencies that children are expected to acquire at a given stage. For instance, the skills described in proficiency frameworks for reading such as the Global Proficiency Framework or one of the several standard setting exercises linked to the Global Alliance to Monitor Learning. Where a study can show not only that the programme raised scores by a certain margin but that a given share of children moved from one recognisable skill level to the next, the results communicate something meaningful regardless of the instrument used. Fluency measures are necessarily language-specific while skill-based levels could serve a parallel function across a wider range of assessments and countries, provided they are based on evidence of what children at each level can actually do. These need not be identical across all settings, but their use requires deep engagement with the measurement process and what results mean for children.

4.2.1. Example: How benchmarks reveal real progress in South Africa

South Africa offers a particularly strong reference point, demonstrating how benchmarks can show whether children improved statistically as well as whether they reached useful levels of skill. Across the last decade, the Early Grade Reading Study (EGRS) has generated longitudinal evidence on the acquisition of foundational reading skills in Setswana and English. Together with the Department of Basic Education's development of reading benchmarks for specific languages, South Africa tells a unique story about how meaningful learning gains should be read, how fluency develops in agglutinative languages, and why benchmark attainment cannot be the sole criterion for judging programme success.

Over three years, the first round of the EGRS showed how a structured pedagogy programme can generate real gains even when most learners remain below comprehension benchmarks. Students gained roughly 8–10 CWPM, improved knowledge of letter sounds and nonword decoding, and saw large reductions in the share scoring zero. Although few reached fluency levels required for comprehension, many children moved from nonreader to emergent or developing reader status. This represents important developmental progress in a language where reliable word reading begins around 20 CWPM and functional comprehension around 35 CWPM. Longitudinal evidence reinforces this picture. Tracking the same learners to Grade 7, Stern et al. (2026) found that early fluency gains persisted seven years later and enabled skills in Setswana and English written comprehension that only emerged later. Only learners who had crossed early fluency thresholds experienced these subsequent comprehension benefits, underscoring that fluency can also function as a gateway skill rather than a standalone metric. This illustrates why benchmark attainment alone is an insufficient measure of outcomes at the child level and programme level, as it fails to capture whether gains persist and translate into higher-order reading competencies over time.

South Africa's experience reinforces several central themes of this paper about how to interpret reading gains. First, progress occurs at multiple developmental levels. For an emergent reader, moving from 0 to 20 CWPM is as important as a shift from 40 to 60 CWPM for a learner approaching proficiency. Second, benchmarks that reflect the language of instruction are essential for making such progress visible. The DBE's commitment to setting benchmarks across all 11 official languages greatly strengthens the interpretation of results. Without Setswana's thresholds of roughly 20 CWPM for early word reading and 35 CWPM for functional comprehension, gains would appear modest when they in fact represent real developmental progress (Ardington et al., 2021; Spaull et al., 2020). Third, fluency gains alone are not sufficient. Although oral reading fluency showed improvement, most learners still fell short of the fluency levels needed for comprehension in either Setswana or English. Finally, South Africa illustrates why distributional shifts matter more than averages. Reductions in share scoring zero and movement from emergent to developing reader bands reveal a richer picture of learning than mean CWPM or effect sizes alone. The goal is not uniformity but transparency, ensuring that literacy reporting reflects how children actually learn and progress toward comprehension.

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

This is not a psychometrics paper, but how we read programme effects ultimately depends on how well the measures fit. Where an assessment is poorly matched to the skills of interest, reported changes will misrepresent true learning gains regardless of how responses are scored or summarised.

In the absence of a widely adopted set of common literacy measures, three issues are important, namely the need for minimum reporting standards regardless of instrument, the prominent but

imperfect role of EGRA, and the need for faster progress towards shared measurement tools that support comparison and clarity.

First, whatever tool is used, researchers should document key properties of the instruments they use. Drawing on Bertling et al. (2023), this includes whether the instrument was piloted in the relevant context; whether score distributions show floor effects (many children scoring zero) or ceiling effects (many scoring perfectly); whether items are internally coherent and show variation across individuals; which skill domains are covered; and how responses are scored and combined. These choices directly affect the reliability of reported scores and, in turn, whether effect estimates support valid inferences about learning gains (American Educational Research Association, 2014).

Second, while the EGRA appears in only around 11 percent of studies in our database, this still makes it the most widely used literacy assessment. It has several features that align with the goals of this paper. In raw form, it reports units that readers can understand directly, such as correct letters identified or words per minute, which makes it easier to see underlying gains alongside standardised effects. The EGRA also has established documentation, and experience across languages in many African countries. Yet at the same time EGRA is not a complete solution. In particular, reading comprehension is weakly measured relative to decoding and fluency. There is also no consistent approach to aggregating subtasks, creating wide researcher discretion and making it harder to compare reported effects. Addressing these issues would improve the consistency and clarity of results in studies that use the EGRA.

Third, alongside improvements in the use of existing tools, there is work ongoing, most notably in relation to the GEEAP, to develop publicly available item banks that link to existing Global Proficiency Frameworks for numeracy and literacy. The aim would not be to impose a single instrument, but allow results from different studies to be placed on a common scale. Where successful, such an approach would improve comparability by reducing dependence on scaling choices tied to individual tests, and make it possible to aggregate evidence across studies without conflating learning differences with instrument artefacts. Such item banks can also lower the costs involved in assessment development, while still allowing local adaptation. Although a common scale has no obvious meaning on its own, this can be addressed by linking scale points to descriptions of what children can typically do, and by using (feasibly common) benchmarks to clarify what counts as real progress. This is a more ambitious agenda which could move the field beyond standardised effect sizes altogether.

5. Summary and conclusion

This paper set out to ask how far standardised effect sizes can be compared across studies when the underlying measures, score distributions and denominators differ. Drawing on data from 197 studies of education interventions in low and middle income countries, the answer is that they often cannot. Dispersion within the reference group, especially floor effects, can inflate standardised effects for

modest raw gains. The choice of reference group for standardisation can shift reported effects in ways that are rarely transparent. And the measurement instrument itself, including its difficulty, construct coverage and scoring rules, strongly conditions the scale on which programme impact is expressed. From assessments in our database, a single additional correct response, or additional word read, can correspond to a standardised effect as small as 0.03 SD or as large as 0.80 SD, depending on the test, the sample and the age of the children assessed.

These distortions matter because they feed into how programmes are ranked and resources are allocated. When standardised effects are used in concepts such as Learning Adjusted Years of Schooling or cost per standard deviation, any inflation in the original estimate is embedded. Current norms risk rewarding large, legible numbers over meaningful learning gains. Journals, funders and evidence synthesis initiatives face strong incentives to highlight single, comparable figures, and researchers in turn have reason to favour instruments, samples and analytical choices which may mechanically increase apparent effects. The cumulative result is a research and policy ecosystem in which the reported size of a programme's impact can drift away from what children can actually do differently.

These findings have specific consequences for how decisions are made. For funders and policymakers, the results highlight the risks of relying on league tables of effect sizes to make allocation decisions. In environments where many children cannot yet read, where instruments suffer from severe floor effects, or where standardisation draws on narrow subsamples, apparently "large" effects may correspond to modest practical change, and vice versa. Donors and education ministries should be cautious about treating single summary metrics as portable across contexts. A programme generating "X SD" in one setting may achieve that figure largely because of properties of the local assessment and sample. Assuming that outcome will apply in other settings risks both overpromising and misallocation. For governments and implementing organisations, the implication is that national discussions of "what works" should focus less on the nominal size of effect sizes and more on where in the distribution programmes move children, and how quickly. A government may reasonably place high value on a programme that reduces the share of nonreaders and moves a sizeable minority to the cusp of fluency, even if average fluency remains below benchmark. An effect size ranking might instead favour a programme with a larger effect concentrated among already better-performing children.

Our call for more information may seem like it runs counter to the push for simplicity and better evidence communication with policymakers. In fact the current convention may be the more demanding one, asking readers to take cross-study comparability on trust when it often does not hold. Reporting raw effects and benchmarks alongside the standardised one lets policymakers and donors judge a study directly. It also disciplines comparisons that advisors may make, making visible the differences in skills assessed, samples, and reference distributions that currently sit hidden behind a common metric, and making it harder to treat as equivalent effects that rest on very different foundations. The score distribution, the choice of denominator and the documentation of

instrument suitability can then strengthen systematic reviews and meta-analyses. There they can be used as study characteristics in a meta-regression or in sensitivity analyses (Deeks et al., 2024). Differences in test design and sample composition may be modelled across studies rather than left implicit in pooled estimates. Identifying the evidence relevant to a given decision then becomes more straightforward, and reliable, not less.

Our analysis has several limitations. This is a descriptive and analytical paper, not a formal systematic review, and the study sample is built to capture variation in measurement and reporting practices rather than to be exhaustive. The detailed fluency analysis rests on a subset of 45 studies that share a common raw metric; the conclusions about raw versus standardised comparisons are clearest in that domain and may or may not transfer exactly to other skill areas. We do not model the relationship between measurement choices and true programme effects econometrically. The analysis demonstrates that the variation is large and consequential, but it does not recover "correct" effect sizes.

As USAID's role in early-grade reading work recedes and the role of EGRA weakens, measurement is likely to fragment further. That makes cross-study synthesis harder, but it also creates an opportunity. If evidence synthesis bodies, journal editors and funders begin to expect raw gains, score distributions and instrument documentation alongside standardised effects, the incentive to report large but opaque effect sizes will fall. If governments invest in language-specific benchmarks and the field develops shared measurement scales, the dependence on study-specific standardisation will weaken. The aim is not to replace one dominant metric with another but to ensure that when programmes are described as effective, those claims rest on clear improvements in children's skills. Policymakers, practitioners and donors want to know which interventions lead to meaningful gains in reading and language. That is the question the evidence should be designed to answer, enabling more children to read with understanding, and to do so sooner.

References

Abadzi, H. (2012). Developing Cross-Language Metrics for Reading Fluency Measurement: Some Issues and Options. Global Partnership for Education Working Paper No. 6. Washington, DC: Global Partnership for Education. https://doi.org/10.1596/26819.

Abadzi, H., & Centanni, M. (2020). *Improving reading outcomes in low-income countries: The role of fluency*. World Bank.

Alsalti, T., Protzko, J., Lakens, D., Elson, M., & Arslan, R. C. (2024). From Ells to Metres: Population norms should supersede sample-local standardisation. https://doi.org/10.31234/osf.io/z34hg.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Angrist, N., Evans, D., Filmer, D., Glennerster, R., Rogers, H., & Sabarwal, S. (2025). How to Improve Education Outcomes Most Efficiently? A Review of the Evidence Using a Unified Metric. *Journal of Development Economics*, 172 (January). https://doi.org/10.1016/j.jdeveco.2024.103382.

Ardington, C., Wills, G., Pretorius, E., Mohohlwane, N., & Menendez, A. (2021). Benchmarking Oral Reading Fluency in the Early Grades in Nguni Languages. *International Journal of Educational Development*, 84, 102433. https://doi.org/10.1016/j.ijedudev.2021.102433.

Asim, S., Chase, R. S., Dar, A., & Schmillen, A. (2017). Improving Learning Outcomes in South Asia: Findings from a Decade of Impact Evaluations. *The World Bank Research Observer*, 32(1), 113–132. https://doi.org/10.1093/wbro/lkw006.

Baguley, T. (2009). Standardized or Simple Effect Size: What Should Be Reported? *British Journal of Psychology*, 100(3), 603–617. https://doi.org/10.1348/000712608X377117.

Baird, M. D., & Pane, J. F. (2019). Translating Standardized Effects of Education Programs into More Interpretable Metrics. *Educational Researcher*, 48(4), 217–228. https://doi.org/10.3102/0013189X19848729.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. *Journal of Economic Perspectives*, 31(4), 73–102. https://doi.org/10.1257/jep.31.4.73.

Bertling, M., Singh, A., & Muralidharan, K. (2023). Psychometric Quality of Measures of Learning Outcomes in Low- and Middle-Income Countries. CGD Working Paper 638. Washington, DC: Center for Global Development.

Cheung, A. C. K., & Slavin, R. E. (2016). How Methodological Features Affect Effect Sizes in Education. *Educational Researcher*, 45(5), 283–292. https://doi.org/10.3102/0013189X16656615.

Conn, K. M. (2017). Identifying Effective Education Interventions in Sub-Saharan Africa: A Meta-Analysis of Impact Evaluations. *Review of Educational Research*, 87(5), 863–898. https://doi.org/10.3102/0034654317712025.

Crawford, M., Raheel, N., Korochkina, M., & Rastle, K. (2024). Inadequate Foundational Decoding Skills Constrain Global Literacy Goals for Pupils in Low- and Middle-Income Countries. *Nature Human Behaviour*, 9, 74–83. https://doi.org/10.1038/s41562-024-02028-x.

Cuijpers, P. (2021). Has the Time Come to Stop Using the "Standardised Mean Difference"? *Clinical Psychology in Europe*, 3(3). https://doi.org/10.32872/cpe.6835.

Deeks, J., Higgins, J., Altman, D., McKenzie, J., Veroniki, A., editor(s). Chapter 10: Analysing data and undertaking meta-analyses (last updated November 2024). In: Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editor(s). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane, 2024. Available from cochrane.org/handbook.

Dowd, A. J., & Bartlett, L. (2019). The Need for Speed: Interrogating the Dominance of Oral Reading Fluency in International Reading Efforts. *Comparative Education Review*, 63(2), 189–212. https://doi.org/10.1086/702612.

Evans, D. K., & Yuan, F. (2022). How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, 44(3), 532–540. https://doi.org/10.3102/01623737221079646.

Ganimian, A. J., & Murnane, R. J. (2016). Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations. *Review of Educational Research*, 86(3), 719–755. https://doi.org/10.3102/0034654315627499.

GEEAP (Global Education Evidence Advisory Panel). (2020). Cost-Effective Approaches to Improve Global Learning. Washington, DC: World Bank.

GEEAP (Global Education Evidence Advisory Panel). (2023a). Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are "Smart Buys" for Improving Learning in Low- and Middle-Income Countries? Washington, DC: World Bank.

GEEAP (Global Education Evidence Advisory Panel). (2023b). GEEAP Database: Global Education Evidence Advisory Panel Study List. Washington, DC: World Bank. https://docs.google.com/spreadsheets/d/1wGMn2g5jUDtkYUKydxqlcEklIGT5wlk4/edit.

GEEAP (Global Education Evidence Advisory Panel). (2023c). Online Appendix: Cost-Effectiveness–Approaches to Improve Global Learning 2023. Washington, DC: World Bank. https://drive.google.com/file/d/1-O8OQRerByxlxsFSZMhDh-58woj8vwxX/view.

Glewwe, P., & Kremer, M. (2006). Schools, Teachers, and Education Outcomes in Developing Countries. In E. A. Hanushek & F. Welch (Eds.), Handbook of the Economics of Education (Vol. 2, pp. 945–1017). Amsterdam: Netherlands. https://doi.org/10.1016/s1574-0692(06)02016-2.

Glewwe, P., & Muralidharan, K. (2016). Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), Handbook of the Economics of Education (Vol. 5, pp. 653–743). Amsterdam: Netherlands. https://doi.org/10.1016/b978-0-444-63459-7.00010-5.

Glewwe, P., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2013). School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. In P. Glewwe (Ed.), Education Policy in Developing Countries (pp. 13–64). Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226078854.003.0002.

Graham, J., & Kelly, S. (2019). How Effective Are Early Grade Reading Interventions? A Review of the Evidence. *Educational Research Review*, 27, 155–175. https://doi.org/10.1016/j.edurev.2019.03.006.

J-PAL (Abdul Latif Jameel Poverty Action Lab). (2017). Roll Call: Getting Children into School. J-PAL Policy Bulletin. Cambridge, MA: MIT.

Jukes, M., Pretorius, E., Schaefer, M., Tjasink, K., Roper, M., Bisgard, J., & Mabhena, N. (2020). Setting reading benchmarks in South Africa. USAID/Khulisa. https://khulisa.com/wp-content/uploads/2020/12/PA00X1NZ.pdf.

Kerwin, J. T., & Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. *Review of Economics and Statistics*, 103(2), 251–264. https://doi.org/10.1162/rest_a_00911.

Kim, Y. G., Lee, H., & Zuilkowski, S. S. (2020). Impact of Literacy Interventions on Reading Skills in Low- and Middle-Income Countries: A Meta-Analysis. *Child Development*, 91(2), 638–660. https://doi.org/10.1111/cdev.13204.

Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking: Methods and Practices (3rd ed.). New York: Springer. https://doi.org/10.1007/978-1-4939-0317-7.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4), 241–253. https://doi.org/10.3102/0013189X20912798.

Krishnaratne, S., White, H., & Carpenter, E. (2013). Quality Education for All Children? What Works in Education in Developing Countries. 3ie Working Paper 20. New Delhi: International Initiative for Impact Evaluation. https://doi.org/10.23846/wp0020.

Masino, S., & Niño-Zarazúa, M. (2016). What Works to Improve the Quality of Student Learning in Developing Countries? *International Journal of Educational Development*, 48, 53–65. https://doi.org/10.1016/j.ijedudev.2015.11.012.

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments. *Review of Educational Research*, 85(3), 353–394. https://doi.org/10.3102/0034654314553127.

Piper, B., & Mugenda, A. (2014). The Primary Math and Reading (PRIMR) Initiative: Endline Impact Evaluation. Research Triangle Park, NC: RTI International, prepared for USAID/Kenya.

Piper, B., Zuilkowski, S. S., & Mugenda, A. (2014). Improving Reading Outcomes in Kenya: First-Year Effects of the PRIMR Initiative. *International Journal of Educational Development*, 37, 11–21. https://doi.org/10.1016/j.ijedudev.2014.02.006.

Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. American Economic Journal: Applied Economics, 6(2), 105–126. https://doi.org/10.1257/app.6.2.105.

Ritchie, E. (2025). FCDO's Best Buys Deserve a Larger Audience. Center for Global Development Blog.

Sandefur, J., Alvares de Azevedo, T., Ju, X., & Le, T. (2023). Phonics and Foreign Aid: Can America Teach the World to Read? CGD Working Paper 668. Washington, DC: Center for Global Development.

Simpson, A. (2017). The Misdirection of Public Policy: Comparing and Combining Standardised Effect Sizes. Journal of Education Policy, 32(4), 450–466. https://doi.org/10.1080/02680939.2017.1280183.

Singh, A. (2015a). How Standard Is the Standard Deviation? A Cautionary Note on Using SDs to Compare Across Impact Evaluations. World Bank Blogs, Development Impact.

Singh, A. (2015b). Private School Effects in Urban and Rural India: Panel Estimates at Primary and Secondary School Ages. *Journal of Development Economics*, 113, 16–32. https://doi.org/10.1016/j.jdeveco.2014.10.004.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, E., Schmidt, T., Jobse, H., Geelen, M., Pastorello, M., & Eyers, J. (2015). Interventions for Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries: A Systematic Review. 3ie Systematic Review 24. London: International Initiative for Impact Evaluation. https://doi.org/10.23846/srs007.

Spaull, N., Pretorius, E., & Mohohlwane, N. (2020). Investigating the Comprehension Iceberg: Developing Empirical Benchmarks for Early-Grade Reading in Agglutinating African Languages. *South African Journal of Childhood Education*, 10(1), a773. https://doi.org/10.4102/sajce.v10i1.773.

Stern, J. M. B., & Piper, B. (2019). Resetting Targets: Examining Large Effect Sizes, Disappointing Benchmark Progress. RTI Press Publication No. OP-0060-1904. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2019.op.0060.1904.

Stern, J. M. B., Jukes, M. C. H., Cilliers, J., Fleisch, B., Taylor, S., & Mohohlwane, N. (2026). Persistence and Emergence of Literacy Skills: Long-Term Impacts of an Effective Early Grade Reading Intervention in South Africa. *Journal of Research on Educational Effectiveness,* 19(1), 1–22. https://doi.org/10.1080/19345747.2024.2417288.

Walls, E. (2025) Impact of USAID withdrawal on global education and skills development – Official development assistance analysis in education and skills development, European Training Foundation, https://data.europa.eu/doi/10.2816/9354807.

Zuilkowski, S. S., Piper, B., Kwayumba, D., & Dubeck, M. (2019). Examining Options for Reading Comprehension Assessment in International Contexts. *Journal of Research in Reading*, 42(3–4), 583–599. https://doi.org/10.1111/1467-9817.12285.

Annex A: What is a standardised effect size?

When researchers report the results of an education intervention, they almost always express them in terms of a **standardised effect size**. It is meant to provide a simple, comparable summary of a programme's impact, regardless of the test used or the scale of its scores.

A standardised effect size typically represents the difference in average performance between the treatment and control groups, expressed in **standard deviations** rather than the original measurement units. The simplest way to calculate it is to divide the raw effect, the difference between the two groups, by the standard deviation of a reference group, most often (but not always) the control group.[27]

For example, at the end of an intervention, children in the treatment group could read 20 words per minute on average and those in the control group 15 words per minute, for a raw programme effect of 5 words per minute. The control group standard deviation is 10 words per minute, so the standardised effect size would be:

$$5 \div 10 = 0.5 \text{ SD}$$

[Raw Effect] ÷ [Std. Deviation] = [Std. Effect Size]

In words, the intervention improved reading by **half of the spread of scores in the control group**. As you can imagine, when that spread is wide, a 0.5 SD gain represents a large absolute difference; when the spread is narrow, the same 0.5 SD implies a much smaller real gain.

The widespread use of standardised effect sizes has grown alongside the rise of randomised control trials in global education. Yet its simplicity conceals three important complications that shape how results should be interpreted.

1. The choice of reference group matters, especially in very-low-literacy environments. Different denominators can produce very different standardised effects for the same underlying gain.
2. The dispersion of scores in that group matters. When the reference distribution is unusually wide or narrow, the same raw effect translates into very different standardised effects.
3. The measurement tool shapes that dispersion. Test content, floor and ceiling effects, aggregation and scoring conventions directly influence how much variation appears in the data and therefore how large the standardised effect size appears.

These three issues underpin the work that follows.

27 Alternatively, 'z-scores' can be used. First, the mean and standard deviation of the scores in a reference group are calculated. Then each individual's raw score is transformed by subtracting the mean and dividing by the standard deviation. This produces a 'z-score' showing how far above or below the mean each observation lies. Analysis proceeds using these 'z-scores'.

Annex B: Study selection and data

B.1. Sample construction

The sample draws on five overlapping sources of studies evaluating educational interventions in low- and middle-income countries:

1. Evans & Yuan (2022), who compiled 234 studies from 11 earlier systematic reviews (Kremer et al., 2013; Krishnaratne & White, 2013; Glewwe et al., 2013; Ganimian & Murnane, 2016; McEwan, 2015; Masino & Niño-Zarazúa, 2016; Glewwe & Muralidharan, 2016; Asim et al., 2017; Snilstveit et al., 2015; Conn, 2017; J-PAL, 2017) supplemented with a literature search covering 2015–2018.
2. The GEEAP database (GEEAP, 2023a, 2023b), a meta-analysis of the cost-effectiveness of education interventions in LMICs (238 studies).
3. Bertling et al. (2023), a review of learning measurement in LMICs (136 studies).
4. Kim et al. (2020), a review of literacy interventions in LMICs (65 studies).
5. A non-exhaustive update search for relevant studies published since 2022, covering:
    a. The African Education Research Database, the American Economic Journal: Applied Economics, the American Economic Journal: Microeconomics, the American Economic Review, Economics of Education Review, Education Economics, the International Journal of Educational Development, the Journal of Development Economics, the Journal of Economic Perspectives, the Journal of Human Resources, and the Quarterly Journal of Economics (51 studies).

From the combined, deduplicated set of candidates we retained peer-reviewed papers that

1. are conducted in a low- or middle-income country (at the time of data collection).
2. measure intervention impacts for children in primary schools
3. report at least one reading, language or literacy outcome; this could be part of a composite index or a language-specific measure
4. use an experimental or quasi-experimental design, including difference-in-differences, regression discontinuity, instrumental variables and propensity score matching, but excluding purely observational studies (our purpose is to examine how the impacts of interventions or policy changes are measured and reported across the literature, not to estimate a mean effect under a certain identification standard).

Applying these criteria yielded a final sample of 171 papers.

B.2. Effect extraction

From each paper's results tables we extracted every reported treatment effect on a language or literacy outcome, recording each estimate in the units used by the authors. Units included standard deviations, raw test-score points, percentage points, or proportions, such as exam pass rates. No conversions or standardisations were applied at the extraction stage. Where a paper reports the same outcome in both SD-normalised and raw-score units, both versions are retained. This produced an effect-level dataset of 2,018 language or literacy measures across the papers. At the same time we recorded the reference group for standardisation, often from table notes but also from elsewhere in the body text, and the assessment(s) used. We use this information descriptively to document reporting patterns.

B.3. Supplementary fluency data from Sandefur et al. (2023)

Nineteen of the 171 papers report oral reading fluency outcomes measured and reported in correct words per minute. To extend coverage of this metric, we supplement the peer-reviewed literature with data from Sandefur et al. (2023), who reanalysed original microdata from USAID-funded Early Grade Reading Assessment (EGRA) evaluations in LMICs. From their sample we retain the 26 evaluations that used an experimental or quasi-experimental design, report both a standardised and a raw (correct words per minute) effect, and do not already appear in our sample.

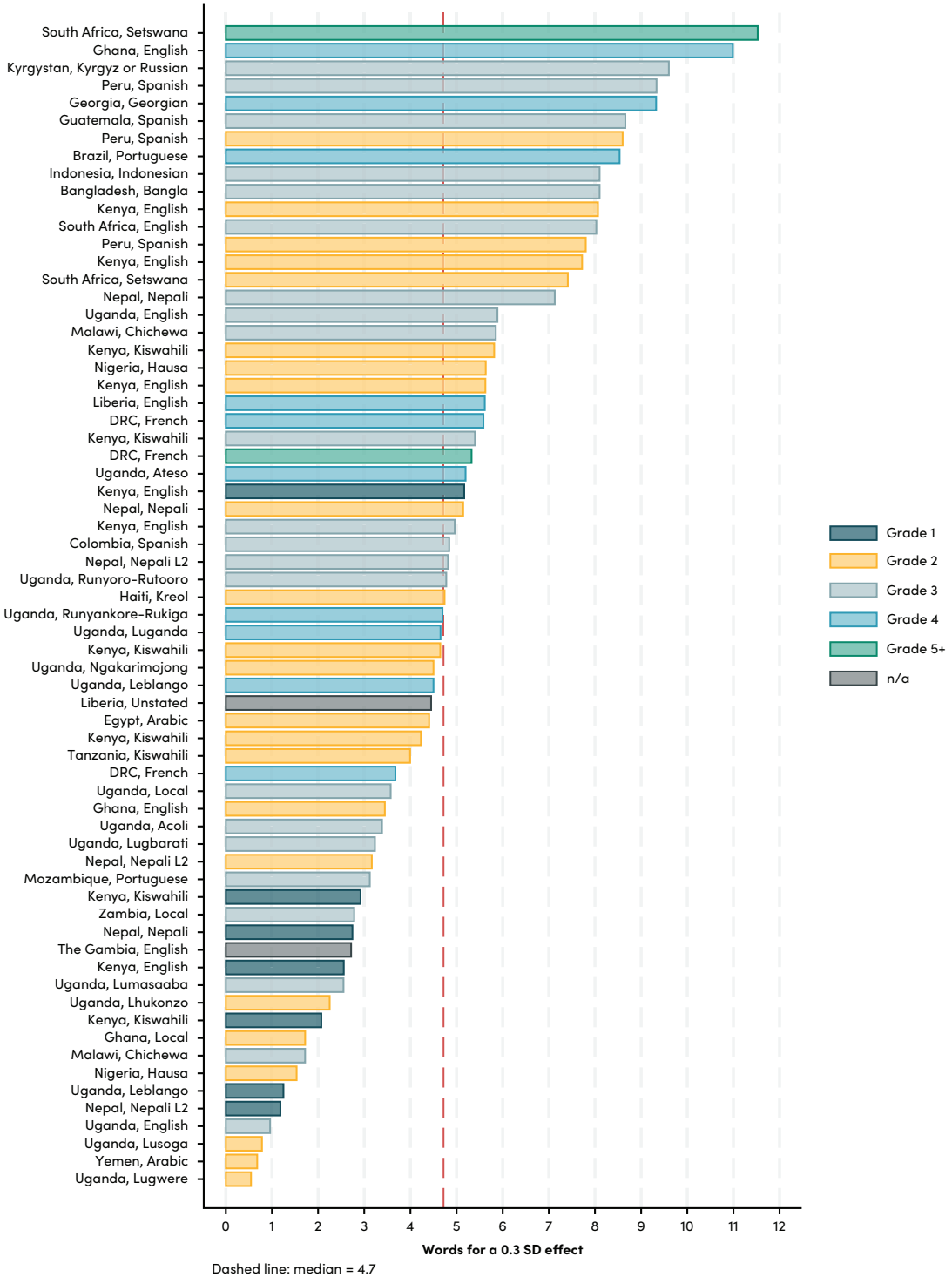
We rely on the summary statistics in Sandefur et al. rather than returning to the original evaluation reports, because the original reports sometimes calculate effect sizes inconsistently and Sandefur et al. applies a common analytical framework to the underlying microdata. These programme evaluations were not published in peer-reviewed journals; we use them only in the reading-fluency analysis because they provide paired raw and standardised effects on a common metric, and we exclude them from our description of reporting trends in the peer-reviewed literature.

B.4. Replication data

We use replication data for 18 studies as illustrative case studies. These studies were not selected at random; we chose them because their microdata allow us to demonstrate the mechanisms discussed in the paper for example, how zero-score shares or reference-group standard deviations, often omitted in published tables, affect the interpretation of standardised effects. We use these files to provide descriptive context and to illustrate denominator sensitivity, not to generate new treatment-effect estimates.

Annex C: Additional words per minute required for a 0.3 SD effect size

FIGURE C1. Additional words per minute required for a 0.3 SD effect size



Notes: This figure inverts the relationship used in Figure 2, providing an alternative perspective. We simply ask how many additional words per minute would a study need to return a uniform 0.3 SD effect size.

Annex D: List of papers included in dataset

1. Abeberese, A. B., Kumler, T. J., & Linden, L. L. (2014). Improving Reading Skills by Encouraging Children to Read in School. Journal of Human Resources, 49(3), 611–633. https://doi.org/10.3368/jhr.49.3.611.

2. Aber, J. L., Tubbs, C., Torrente, C., Halpin, P. F., Johnston, B., Starkey, L., Shivshanker, A., Annan, J., Seidman, E., & Wolf, S. (2017). Promoting children's learning and development in conflict-affected countries: Testing change process in the Democratic Republic of the Congo. Development and Psychopathology, 29(1), 53–67. https://doi.org/10.1017/S0954579416001139.

3. Abrami, P. C., Wade, C. A., Lysenko, L., Marsh, J., & Gioko, A. (2016). Using educational technology to develop early literacy skills in Sub-Saharan Africa. Education and Information Technologies, 21(4), 945–964. https://doi.org/10.1007/s10639-014-9362-4.

4. Abrigo, M. R. & Francisco, K. A. (2024). Compulsory kindergarten education and early teenage literacy in the Philippines. International Journal of Educational Development, 109, 103087. https://doi.org/10.1016/j.ijedudev.2024.103087.

5. Adrogue, C. & Orlicki, M. E. (2013). Do In-School Feeding Programs Have an Impact on Academic Performance and Dropouts? The Case of Public Schools in Argentina. Education Policy Analysis Archives, 21, 50. https://doi.org/10.14507/epaa.v21n50.2013.

6. Adukia, A. (2017). Sanitation and Education. American Economic Journal: Applied Economics, 9(2), 23–59. https://doi.org/10.1257/app.20150083.

7. Alvarez-Marinelli, H., Blanco, M., Lara-Alecio, R., Irby, B. J., Tong, F., Stanley, K., & Fan, Y. (2016). Computer assisted English language learning in Costa Rican elementary schools: an experimental study. Computer Assisted Language Learning, 29(1), 103–126. https://doi.org/10.1080/09588221.2014.903977.

8. Anand, P., Mizala, A., & Repetto, A. (2009). Using school scholarships to estimate the effect of private education on the academic achievement of low-income students in Chile. Economics of Education Review, 28(3), 370–381. https://doi.org/10.1016/j.econedurev.2008.03.005.

9. Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets. American Economic Review, 107(6), 1535–1563. https://doi.org/10.1257/aer.20140774.

10. Andrabi, T., Bau, N., Das, J., Karachiwalla, N., & Ijaz Khwaja, A. (2024). Crowding in Private Quality: The Equilibrium Effects of Public Spending in Education. The Quarterly Journal of Economics, 139(4), 2525–2577. https://doi.org/10.1093/qje/qjae014.

11. Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. American Economic Review, 96(3), 847–862. https://doi.org/10.1257/aer.96.3.847.

12. Aturupane, H., Glewwe, P., Ravina, R., Sonnadara, U., & Wisniewski, S. (2014). An Assessment of the Impacts of Sri Lanka's Programme for School Improvement and School Report Card Programme on Students' Academic Progress. The Journal of Development Studies, 50(12), 1647–1669. https://doi.org/10.1080/00220388.2014.936396.

13. Aturupane, H., Glewwe, P., Utsumi, T., Wisniewski, S., & Shojo, M. (2022). The impact of Sri Lanka's school-based management programme on teachers' pedagogical practices and student learning: evidence from a randomised controlled trial. Journal of Development Effectiveness, 14(4), 285–305. https://doi.org/10.1080/19439342.2022.2029540.

14. Aurino, E., Gelli, A., Adamba, C., Osei-Akoto, I., & Alderman, H. (2023). Food for Thought? Journal of Human Resources, 58(1), 74–111. https://doi.org/10.3368/jhr.58.3.1019-10515R1.

15. Bai, Y., Tang, B., Wang, B., Mo, D., Zhang, L., Rozelle, S., Auden, E., & Mandell, B. (2023). Impact of online computer assisted learning on education: Experimental evidence from economically vulnerable areas of China. Economics of Education Review, 94, 102385. https://doi.org/10.1016/j.econedurev.2023.102385.

16. Baird, S., McIntosh, C., & Ozler, B. (2011). Cash or Condition? Evidence from a Cash Transfer Experiment. The Quarterly Journal of Economics, 126(4), 1709–1753. https://doi.org/10.1093/qje/qjr032.

17. Baird, S., McIntosh, C., & Özler, B. (2019). When the money runs out: Do cash transfers have sustained effects on human capital accumulation?. Journal of Development Economics, 140, 169–185. https://doi.org/10.1016/j.jdeveco.2019.04.004.

18. Bando, R., Gallego, F., Gertler, P., & Romero Fonseca, D. (2017). Books or laptops? The effect of shifting from printed to digital delivery of educational content on learning. Economics of Education Review, 61, 162–173. https://doi.org/10.1016/j.econedurev.2017.07.005.

19. Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. The Quarterly Journal of Economics, 122(3), 1235–1264. https://doi.org/10.1162/qjec.122.3.1235.

20. Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. American Economic Journal: Economic Policy, 2(1), 1–30. https://doi.org/10.1257/pol.2.1.1.

21. Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives, 31(4), 73–102. https://doi.org/10.1257/jep.31.4.73.

22. Banerji, R., Berry, J., & Shotland, M. (2017). The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India. American Economic Journal: Applied Economics, 9(4), 303–337. https://doi.org/10.1257/app.20150390.

23. Barrera-Osorio, F. & Raju, D. (2017). Teacher performance pay: Experimental evidence from Pakistan. Journal of Public Economics, 148, 75–91. https://doi.org/10.1016/j.jpubeco.2017.02.001.

24. Barrera-Osorio, F., Galbert, P. D., Habyarimana, J., & Sabarwal, S. (2020). The Impact of Public-Private Partnerships on Private School Performance: Evidence from a Randomized Controlled Trial in Uganda. Economic Development and Cultural Change, 68(2), 429–469. https://doi.org/10.1086/701229.

25. Barrera-Osorio, F., Cilliers, J., Cloutier, M., & Filmer, D. (2022). Heterogenous teacher effects of two incentive schemes: Evidence from a low-income country. Journal of Development Economics, 156, 102820. https://doi.org/10.1016/j.jdeveco.2022.102820.

26. Barrera-Osorio, F., Blakeslee, D. S., Hoover, M., Linden, L., Raju, D., & Ryan, S. P. (2022). Delivering Education to the Underserved through a Public-Private Partnership Program in Pakistan. The Review of Economics and Statistics, 104(3), 399–416. https://doi.org/10.1162/rest_a_01002.

27. Beasley, E. & Huillery, E. (2016). Willing but Unable? Short-Term Experimental Evidence on Parent Empowerment and School Quality. The World Bank Economic Review, lhv064. https://doi.org/10.1093/wber/lhv064.

28. Bellei, C. (2009). Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile. Economics of Education Review, 28(5), 629–640. https://doi.org/10.1016/j.econedurev.2009.01.008.

29. Berlinski, S., Galiani, S., & Gertler, P. (2009). The effect of pre-primary education on primary school performance. Journal of Public Economics, 93(1–2), 219–234. https://doi.org/10.1016/j.jpubeco.2008.09.002.

30. Berry, J. (2015). Child Control in Education Decisions. Journal of Human Resources, 50(4), 1051–1080. https://doi.org/10.3368/jhr.50.4.1051.

31. Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., & Cruz-Aguayo, Y. (2015). One Laptop per Child at Home: Short-Term Impacts from a Randomized Experiment in Peru. American Economic Journal: Applied Economics, 7(2), 53–80. https://doi.org/10.1257/app.20130267.

32. Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2018). Experimental evidence on scaling up education reforms in Kenya. Journal of Public Economics, 168, 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007.

33. Brown, C., Kaur, S., Kingdon, G., & Schofield, H. (2025). Cognitive Endurance as Human Capital. The Quarterly Journal of Economics, 140(2), 943–1002. https://doi.org/10.1093/qje/qjae043.

34. Brunette, T., Piper, B., Jordan, R., King, S., & Nabacwa, R. (2019). The Impact of Mother Tongue Reading Instruction in Twelve Ugandan Languages and the Role of Language Complexity, Socioeconomic Factors, and Program Implementation. Comparative Education Review, 63(4), 591–612. https://doi.org/10.1086/705426.

35. Bruns, B., Costa, L., & Cunha, N. (2018). Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil?. Economics of Education Review, 64, 214–250. https://doi.org/10.1016/j.econedurev.2018.03.003.

36. Buhl-Wiggers, J., Kerwin, J. T., Muñoz-Morales, J., Smith, J., & Thornton, R. (2024). Some children left behind: Variation in the effects of an educational intervention. Journal of Econometrics, 243(1–2), 105256. https://doi.org/10.1016/j.jeconom.2021.12.010.

37. Burde, D. & Linden, L. L. (2013). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. American Economic Journal: Applied Economics, 5(3), 27–40. https://doi.org/10.1257/app.5.3.27.

38. Cardim, J., Molina-Millán, T., & Vicente, P. C. (2023). Can technology improve the classroom experience in primary education? An African experiment on a worldwide program. Journal of Development Economics, 164, 103145. https://doi.org/10.1016/j.jdeveco.2023.103145.

39. Carneiro, P., Koussihouèdé, O., Lahire, N., Meghir, C., & Mommaerts, C. (2020). School Grants and Education Quality: Experimental Evidence from Senegal. Economica, 87(345), 28–51. https://doi.org/10.1111/ecca.12302.

40. Carneiro, P., Cruz-Aguayo, Y., Salvati, F., & Schady, N. (2025). The Effect of Classroom Rank on Learning throughout Elementary School: Experimental Evidence from Ecuador. Journal of Labor Economics, 43(2), 433–466. https://doi.org/10.1086/727515.

41. Chen, X., Liu, C., Zhang, L., Shi, Y., & Rozelle, S. (2010). Does taking one step back get you two steps forward? Grade retention and school performance in poor areas in rural China. International Journal of Educational Development, 30(6), 544–559. https://doi.org/10.1016/j.ijedudev.2009.12.002.

42. Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2020). How to Improve Teaching Practice? Journal of Human Resources, 55(3), 926–962. https://doi.org/10.3368/jhr.55.3.0618-9538r1.

43. Cilliers, J., Fleisch, B., Kotze, J., Mohohlwane, N., Taylor, S., & Thulare, T. (2022). Can virtual replace in-person coaching? Experimental evidence on teacher professional development and student learning. Journal of Development Economics, 155, 102815. https://doi.org/10.1016/j.jdeveco.2021.102815.

44. Clarke, S. E., Rouhani, S., Diarra, S., Saye, R., Bamadio, M., Jones, R., Traore, D., Traore, K., Jukes, M. C., Thuilliez, J., Brooker, S., Roschnik, N., & Sacko, M. (2017). Impact of a malaria intervention package in schools on Plasmodium infection, anaemia and cognitive function in schoolchildren in Mali: a pragmatic cluster-randomised trial. BMJ Global Health, 2(2), e000182. https://doi.org/10.1136/bmjgh-2016-000182.

45. Contreras, D. & Rau, T. (2012). Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile. Economic Development and Cultural Change, 61(1), 219–246. https://doi.org/10.1086/666955.

46. Correa, J. A., Parro, F., & Reyes, L. (2014). The Effects of Vouchers on School Results: Evidence from Chile's Targeted Voucher Program. Journal of Human Capital, 8(4), 351–398. https://doi.org/10.1086/679282.

47. Crawford, M., Rutkowski, D., & Rutkowski, L. (2023). Improving reading abilities, attitudes and practices: A home-based intervention of supplementary texts for young readers in Cambodia. International Journal of Educational Development, 103, 102906. https://doi.org/10.1016/j.ijedudev.2023.102906.

48. Crawfurd, L., Evans, D. K., Hares, S., & Sandefur, J. (2023). Live tutoring calls did not improve learning during the COVID-19 pandemic in Sierra Leone. Journal of Development Economics, 164, 103114. https://doi.org/10.1016/j.jdeveco.2023.103114.

49. Crawfurd, L. & Alam, A. (2023). Contracting out schools at scale: evidence from Pakistan. Education Economics, 31(5), 555–571. https://doi.org/10.1080/09645292.2022.2113859.

50. Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2017). Technology and Child Development: Evidence from the One Laptop per Child Program. American Economic Journal: Applied Economics, 9(3), 295–320. https://doi.org/10.1257/app.20150385.

51. Croke, K. & Atun, R. (2019). The long run impact of early childhood deworming on numeracy and literacy: Evidence from Uganda. PLOS Neglected Tropical Diseases, 13(1), e0007085. https://doi.org/10.1371/journal.pntd.0007085.

52. Cueto, S. & Chinen, M. (2008). Educational impact of a school breakfast programme in rural Peru. International Journal of Educational Development, 28(2), 132–148. https://doi.org/10.1016/j.ijedudev.2007.02.007.

53. Cummins, J. R. (2017). Heterogeneous treatment effects in the low track: Revisiting the Kenyan primary school experiment. Economics of Education Review, 56, 40–51. https://doi.org/10.1016/j.econedurev.2016.11.006.

54. Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School Inputs, Household Substitution, and Test Scores. American Economic Journal: Applied Economics, 5(2), 29–57. https://doi.org/10.1257/app.5.2.29.

55. de Hoyos, R., Garcia-Moreno, V. A., & Patrinos, H. A. (2017). The impact of an accountability intervention with diagnostic feedback: Evidence from Mexico. Economics of Education Review, 58, 123–140. https://doi.org/10.1016/j.econedurev.2017.03.007.

56. de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. (2018). Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia. The Quarterly Journal of Economics, 133(2), 993–1039. https://doi.org/10.1093/qje/qjx040.

57. Delavallade, C., Griffith, A., & Thornton, R. (2021). Effects of a Multi-Faceted Education Program on Enrollment, Learning and Gender Equity: Evidence from India. The World Bank Economic Review, 35(4), 950–968. https://doi.org/10.1093/wber/lhaa025.

58. Dinkelman, T. & Martínez A., C. (2014). Investing in Schooling In Chile: The Role of Information about Financial Aid for Higher Education. The Review of Economics and Statistics, 96(2), 244–257. https://doi.org/10.1162/rest_a_00384.

59. Dixon, P., Schagen, I., & Seedhouse, P. (2011). The impact of an intervention on children's reading and spelling ability in low-income schools in India. School Effectiveness and School Improvement, 22(4), 461–482. https://doi.org/10.1080/09243453.2011.625125.

60. Duflo, E., Dupas, P., & Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. American Economic Review, 101(5), 1739–1774. https://doi.org/10.1257/aer.101.5.1739.

61. Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives Work: Getting Teachers to Come to School. American Economic Review, 102(4), 1241–1278. https://doi.org/10.1257/aer.102.4.1241.

62. Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. Journal of Public Economics, 123, 92–110. https://doi.org/10.1016/j.jpubeco.2014.11.008.

63. Duflo, A., Kiessel, J., & Lucas, A. M. (2024). Experimental Evidence on Four Policies to Increase Learning at Scale. The Economic Journal, 134(661), 1985–2008. https://doi.org/10.1093/ej/ueae003.

64. Ebenezer, R., Gunawardena, K., Kumarendran, B., Pathmeswaran, A., Jukes, M. C. H., Drake, L. J., & de Silva, N. (2013). Cluster-randomised trial of the impact of school-based deworming and iron supplementation on the cognitive abilities of schoolchildren in Sri Lanka's plantation sector. Tropical Medicine & International Health, 18(8), 942–951. https://doi.org/10.1111/tmi.12128.

65. Eble, A., Frost, C., Camara, A., Bouy, B., Bah, M., Sivaraman, M., Hsieh, P. J., Jayanty, C., Brady, T., Gawron, P., Vansteelandt, S., Boone, P., & Elbourne, D. (2021). How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia. Journal of Development Economics, 148, 102539. https://doi.org/10.1016/j.jdeveco.2020.102539.

66. Edmonds, E. V. & Shrestha, M. (2014). You get what you pay for: Schooling incentives and child labor. Journal of Development Economics, 111, 196–211. https://doi.org/10.1016/j.jdeveco.2014.09.005.

67. Fazzio, I., Eble, A., Lumsdaine, R. L., Boone, P., Bouy, B., Hsieh, P.-T. J., Jayanty, C., Johnson, S., & Silva, A. F. (2021). Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau. Journal of Public Economics, 199, 104385. https://doi.org/10.1016/j.jpubeco.2021.104385.

68. Fernando, D., De Silva, D., Carter, R., Mendis, K. N., & Wickremasinghe, R. (2006). A Randomized, Double-Blind, Placebo-Controlled, Clinical Trial of the Impact of Malaria Prevention on the Educational Attainment of School Children. The American Journal of Tropical Medicine and Hygiene, 74(3), 386–393. https://doi.org/10.4269/ajtmh.2006.74.386.

69. Filmer, D., Habyarimana, J., & Sabarwal, S. (2025). Teacher Performance-Based Incentives and Learning Inequality. Journal of Human Resources, 60(3), 812–856. https://doi.org/10.3368/jhr.1120-11313r2.

70. Galiani, S., Gertler, P., & Schargrodsky, E. (2008). School decentralization: Helping the good get better, but leaving the poor behind. Journal of Public Economics, 92(10–11), 2106–2120. https://doi.org/10.1016/j.jpubeco.2008.05.004.

71. Ganimian, A. J., Muralidharan, K., & Walters, C. R. (2024). Augmenting State Capacity for Child Development: Experimental Evidence from India. Journal of Political Economy, 132(5), 1565–1602. https://doi.org/10.1086/728109.

72. Gao, Q., Wang, H., Mo, D., Shi, Y., Kenny, K., & Rozelle, S. (2018). Can reading programs improve reading skills and academic performance in rural China?. China Economic Review, 52, 111–125. https://doi.org/10.1016/j.chieco.2018.07.001.

73. Garcia, S. & Hill, J. (2010). Impact of conditional cash transfers on children's school achievement: evidence from Colombia. Journal of Development Effectiveness, 2(1), 117–137. https://doi.org/10.1080/19439341003628681.

74. Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. Journal of Development Economics, 74(1), 251–268. https://doi.org/10.1016/j.jdeveco.2003.12.010.

75. Glewwe, P., Kremer, M., & Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. American Economic Journal: Applied Economics, 1(1), 112–135. https://doi.org/10.1257/app.1.1.112.

76. Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher Incentives. American Economic Journal: Applied Economics, 2(3), 205–227. https://doi.org/10.1257/app.2.3.205.

77. Glewwe, P. & Maïga, E. W. H. (2011). The impacts of school management reforms in Madagascar: do the impacts vary by teacher type?. Journal of Development Effectiveness, 3(4), 435–469. https://doi.org/10.1080/19439342.2011.604729.

78. Glewwe, P., Park, A., & Zhao, M. (2016). A better vision for development: Eyeglasses and academic performance in rural primary schools in China. Journal of Development Economics, 122, 170–182. https://doi.org/10.1016/j.jdeveco.2016.05.007.

79. Gomes, M. & Hirata, G. (2025). Developing reading fluency in primary school: evidence from an RCT. Education Economics, 33(3), 355–371. https://doi.org/10.1080/09645292.2024.2322535.

80. Halliday, K. E., Okello, G., Turner, E. L., Njagi, K., Mcharo, C., Kengo, J., Allen, E., Dubeck, M. M., Jukes, M. C. H., & Brooker, S. J. (2014). Impact of Intermittent Screening and Treatment for Malaria among School Children in Kenya: A Cluster Randomised Trial. PLoS Medicine, 11(1), e1001594. https://doi.org/10.1371/journal.pmed.1001594.

81. Hasan, A., Jung, H., Kinnell, A., Maika, A., Nakajima, N., & Pradhan, M. (2021). Contrasting Experiences: Understanding the Longer-Term Impact of Improving Access to Pre-Primary Education in Rural Indonesia. Journal of Research on Educational Effectiveness, 14(1), 28–56. https://doi.org/10.1080/19345747.2020.1839989.

82. Hassan, H., Islam, A., Siddique, A., & Wang, L. C. (2024). Telementoring and Homeschooling During School Closures: a Randomised Experiment in Rural Bangladesh. The Economic Journal, 134(662), 2418–2438. https://doi.org/10.1093/ej/ueae014.

83. Heinrich, C. J. (2007). Demand and Supply-Side Determinants of Conditional Cash Transfer Program Effectiveness. World Development, 35(1), 121–143. https://doi.org/10.1016/j.worlddev.2006.09.009.

84. Hoekstra, M., Mouganie, P., & Wang, Y. (2018). Peer Quality and the Academic Benefits to Attending Better Schools. Journal of Labor Economics, 36(4), 841–884. https://doi.org/10.1086/697465.

85. Hsieh, C. & Urquiola, M. (2006). The effects of generalized school choice on achievement and stratification: Evidence from Chile's voucher program. Journal of Public Economics, 90(8–9), 1477–1503. https://doi.org/10.1016/j.jpubeco.2005.11.002.

86. Hulett, J. L., Weiss, R. E., Bwibo, N. O., Galal, O. M., Drorbaugh, N., & Neumann, C. G. (2014). Animal source foods have a positive impact on the primary school test scores of Kenyan schoolchildren in a cluster-randomised, controlled feeding intervention trial. British Journal of Nutrition, 111(5), 875–886. https://doi.org/10.1017/s0007114513003310.

87. Islam, A. (2019). Parent–teacher meetings and student outcomes: Evidence from a developing country. European Economic Review, 111, 273–304. https://doi.org/10.1016/j.euroecorev.2018.09.008.

88. Islam, M. J., Hossain, M., & Haque, S. T. (2025). The cost of discipline? Exploring the impact of corporal punishment on children's foundational learning skills in Bangladesh. International Journal of Educational Development, 117, 103360. https://doi.org/10.1016/j.ijedudev.2025.103360.

89. Jayachandran, S. (2014). Incentives to teach badly: After-school tutoring in developing countries. Journal of Development Economics, 108, 190–205. https://doi.org/10.1016/j.jdeveco.2014.02.008.

90. Jere-Folotiya, J., Chansa-Kabali, T., Munachaka, J. C., Sampa, F., Yalukanda, C., Westerholm, J., Richardson, U., Serpell, R., & Lyytinen, H. (2014). The effect of using a mobile literacy game to improve literacy levels of grade one students in Zambian schools. Educational Technology Research and Development, 62(4), 417–436. https://doi.org/10.1007/s11423-014-9342-9.

91. Jimenez, E. & Sawada, Y. (1999). Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program. The World Bank Economic Review, 13(3), 415–441. https://doi.org/10.1093/wber/13.3.415.

92. Johnston, J. & Ksoll, C. (2022). Effectiveness of interactive satellite-transmitted instruction: Experimental evidence from Ghanaian primary schools. Economics of Education Review, 91, 102315. https://doi.org/10.1016/j.econedurev.2022.102315.

93. Jukes, M. C. H., Pinder, M., Grigorenko, E. L., Smith, H. B., Walraven, G., Bariau, E. M., Sternberg, R. J., Drake, L. J., Milligan, P., Cheung, Y. B., Greenwood, B. M., & Bundy, D. A. P. (2006). Long-Term Impact of Malaria Chemoprophylaxis on Cognitive Abilities and Educational Attainment: Follow-Up of a Controlled Trial. PLoS Clinical Trials, 1(4), e19. https://doi.org/10.1371/journal.pctr.0010019.

94. Jukes, M. C. H., Turner, E. L., Dubeck, M. M., Halliday, K. E., Inyega, H. N., Wolf, S., Zuilkowski, S. S., & Brooker, S. J. (2017). Improving Literacy Instruction in Kenya Through Teacher Professional Development and Text Messages Support: A Cluster Randomized Trial. Journal of Research on Educational Effectiveness, 10(3), 449–481. https://doi.org/10.1080/19345747.2016.1221487.

95. Karlan, D. & Linden, L. L. (2025). Loose knots: Strong versus weak commitments to save for education in Uganda. Journal of Development Economics, 174, 103444. https://doi.org/10.1016/j.jdeveco.2024.103444.

96. Kazianga, H., Levy, D., Linden, L. L., & Sloan, M. (2013). The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. American Economic Journal: Applied Economics, 5(3), 41–62. https://doi.org/10.1257/app.5.3.41.

97. Kerwin, J. T. & Thornton, R. L. (2021). Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. The Review of Economics and Statistics, 103(2), 251–264. https://doi.org/10.1162/rest_a_00911.

98. Khattri, N., Ling, C., & Jha, S. (2012). The effects of school-based management in the Philippines: an initial assessment using administrative data. Journal of Development Effectiveness, 4(2), 277–295. https://doi.org/10.1080/19439342.2012.692389.

99. Koima, J. (2024). School electrification and academic outcomes in rural Kenya. Journal of Development Economics, 166, 103178. https://doi.org/10.1016/j.jdeveco.2023.103178.

100. Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. Review of Economics and Statistics, 91(3), 437–456. https://doi.org/10.1162/rest.91.3.437.

101. Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. Economics of Education Review, 47, 34–48. https://doi.org/10.1016/j.econedurev.2015.03.005.

102. Lai, F., Zhang, L., Qu, Q., Hu, X., Shi, Y., Boswell, M., & Rozelle, S. (2015). Teaching the Language of Wider Communication, Minority Students, and Overall Educational Performance: Evidence from a Randomized Experiment in Qinghai Province, China. Economic Development and Cultural Change, 63(4), 753–776. https://doi.org/10.1086/681233.

103. Lai, F., Zhang, L., Bai, Y., Liu, C., Shi, Y., Chang, F., & Rozelle, S. (2016). More is not always better: evidence from a randomised experiment of computer-assisted learning in rural minority schools in Qinghai. Journal of Development Effectiveness, 8(4), 449–472. https://doi.org/10.1080/19439342.2016.1220412.

104. Laitin, D. D., Ramachandran, R., & Walter, S. L. (2019). The Legacy of Colonial Language Policies and Their Impact on Student Learning: Evidence from an Experimental Program in Cameroon. Economic Development and Cultural Change, 68(1), 239–272. https://doi.org/10.1086/700617.

105. Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., & Mann, V. (2013). The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support. PLoS ONE, 8(7), e65775. https://doi.org/10.1371/journal.pone.0065775.

106. Lara, B., Mizala, A., & Repetto, A. (2011). The Effectiveness of Private Voucher Education. Educational Evaluation and Policy Analysis, 33(2), 119–137. https://doi.org/10.3102/0162373711402990.

107. Lassibille, G., Tan, J.-P., Jesse, C., & Van Nguyen, T. (2010). Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions. The World Bank Economic Review, 24(2), 303–329. https://doi.org/10.1093/wber/lhq009.

108. Lauterbach, S., Crawfurd, L., Kirezi, J. C., Nsabimana, A., & Peeraer, J. (2025). Improving school leadership in Rwanda. Journal of Development Economics, 177, 103545. https://doi.org/10.1016/j.jdeveco.2025.103545.

109. Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2021). Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools. American Economic Review, 111(7), 2213–2246. https://doi.org/10.1257/aer.20191972.

110. Leme, M. C., Louzano, P., Ponczek, V., & Souza, A. P. (2012). The impact of structured teaching methods on the quality of education in Brazil. Economics of Education Review, 31(5), 850–860. https://doi.org/10.1016/j.econedurev.2012.05.008.

111. Li, T., Han, L., Zhang, L., & Rozelle, S. (2014). Encouraging classroom peer interactions: Evidence from Chinese migrant schools. Journal of Public Economics, 111, 29–45. https://doi.org/10.1016/j.jpubeco.2013.12.014.

112. Lucas, A. M., McEwan, P. J., Ngware, M., & Oketch, M. (2014). Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda. Journal of Policy Analysis and Management, 33(4), 950–976. https://doi.org/10.1002/pam.21782.

113. Malamud, O. & Pop-Eleches, C. (2011). Home Computer Use and the Development of Human Capital. The Quarterly Journal of Economics, 126(2), 987–1027. https://doi.org/10.1093/qje/qjr008.

114. Marinelli, H. A., Berlinski, S., & Busso, M. (2024). Remedial Education. Journal of Human Resources, 59(1), 141–174. https://doi.org/10.3368/jhr.0320-10801R2.

115. Marshall, J. H. & Bunly, S. (2017). School grants and school performance in rural Cambodia. Journal of Development Effectiveness, 9(3), 305–328. https://doi.org/10.1080/19439342.2017.1338306.

116. Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. The Quarterly Journal of Economics, 134(3), 1627–1673. https://doi.org/10.1093/qje/qjz010.

117. Mbiti, I., Romero, M., & Schipper, Y. (2023). Designing Effective Teacher Performance Pay Programs: Experimental Evidence from Tanzania. The Economic Journal, 133(653), 1968–2000. https://doi.org/10.1093/ej/uead010.

118. McEwan, P. J. (1998). The effectiveness of multigrade schools in Colombia. International Journal of Educational Development, 18(6), 435–452. https://doi.org/10.1016/s0738-0593(98)00023-6.

119. McEwan, P. J. (2013). The impact of Chile's school feeding program on education outcomes. Economics of Education Review, 32, 122–139. https://doi.org/10.1016/j.econedurev.2012.08.006.

120. McManus, J., Rudasingwa, M., Nijhof, E., Mokobi, K., Deme, S. F., Kiawoin, J., Garay, F. A., Mwai, L., & Pignon, C. (2025). Improving learning outcomes for out-of-school children: evidence from a randomized evaluation of an accelerated learning program in Liberia. Education Economics, 33(1), 19–38. https://doi.org/10.1080/09645292.2024.2410773.

121. Menendez, A. & Haugan, G. (2025). Improving early grade reading performance in Nepal: differences between Nepali and non-Nepali students. Education Economics, 1–17. https://doi.org/10.1080/09645292.2025.2518376.

122. Menezes-Filho, N. & Pazello, E. (2007). Do teachers' wages matter for proficiency? Evidence from a funding reform in Brazil. Economics of Education Review, 26(6), 660–672. https://doi.org/10.1016/j.econedurev.2007.08.003.

123. Miguel, E. & Kremer, M. (2004). Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities. Econometrica, 72(1), 159–217. https://doi.org/10.1111/j.1468-0262.2004.00481.x.

124. Mo, D., Swinnen, J., Zhang, L., Yi, H., Qu, Q., Boswell, M., & Rozelle, S. (2013). Can One-to-One Computing Narrow the Digital Divide and the Educational Gap in China? The Case of Beijing Migrant Schools. World Development, 46, 14–29. https://doi.org/10.1016/j.worlddev.2012.12.019.

125. Mo, D., Bai, Y., Shi, Y., Abbey, C., Zhang, L., Rozelle, S., & Loyalka, P. (2020). Institutions, implementation, and program effectiveness: Evidence from a randomized evaluation of computer-assisted learning in rural China. Journal of Development Economics, 146, 102487. https://doi.org/10.1016/j.jdeveco.2020.102487.

126. Morabito, C., Van de gaer, D., Figueroa, J. L., & Vandenbroeck, M. (2018). Effects of high versus low-quality preschool education: A longitudinal study in Mauritius. Economics of Education Review, 65, 126–137. https://doi.org/10.1016/j.econedurev.2018.06.006.

127. Moussa, W. & Koester, E. (2022). Effects of Story Read-Aloud Lessons on Literacy Development in the Early Grades: Experimental Evidence From Nigeria. Reading Research Quarterly, 57(2), 587–607. https://doi.org/10.1002/rrq.427.

128. Muralidharan, K. & Sundararaman, V. (2010). The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India. The Economic Journal, 120(546), F187–F203. https://doi.org/10.1111/j.1468-0297.2010.02373.x.

129. Muralidharan, K. & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. Journal of Political Economy, 119(1), 39–77. https://doi.org/10.1086/659655.

130. Muralidharan, K. & Sundararaman, V. (2015). The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India. The Quarterly Journal of Economics, 130(3), 1011–1066. https://doi.org/10.1093/qje/qjv013.

131. Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. American Economic Review, 109(4), 1426–1460. https://doi.org/10.1257/aer.20171112.

132. Naseer, M. F., Patnam, M., & Raza, R. R. (2010). Transforming public schools: Impact of the CRI program on child learning in Pakistan. Economics of Education Review, 29(4), 669–683. https://doi.org/10.1016/j.econedurev.2009.12.001.

133. Nonoyama-Tarumi, Y. & Bredenberg, K. (2009). Impact of school readiness program interventions on children's learning in Cambodia. International Journal of Educational Development, 29(1), 39–45. https://doi.org/10.1016/j.ijedudev.2008.07.003.

134. Olken, B. A., Onishi, J., & Wong, S. (2014). Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia. American Economic Journal: Applied Economics, 6(4), 1–34. https://doi.org/10.1257/app.6.4.1.

135. Pallante, D. H. & Kim, Y. (2013). The effect of a multicomponent literacy instruction model on literacy growth for kindergartners and first-grade students in Chile. International Journal of Psychology, 48(5), 747–761. https://doi.org/10.1080/00207594.2012.719628.

136. Pandey, P., Goyal, S., & Sundararaman, V. (2009). Community participation in public schools: impact of information campaigns in three Indian states. Education Economics, 17(3), 355–375. https://doi.org/10.1080/09645290903157484.

137. Piper, B., Zuilkowski, S. S., Kwayumba, D., & Strigel, C. (2016). Does technology improve reading outcomes? Comparing the effectiveness and cost-effectiveness of ICT interventions for early grade reading in Kenya. International Journal of Educational Development, 49, 204–214. https://doi.org/10.1016/j.ijedudev.2016.03.006.

138. Piper, B., Zuilkowski, S. S., & Ong'ele, S. (2016). Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya. Comparative Education Review, 60(4), 776–807. https://doi.org/10.1086/688493.

139. Piper, B., Destefano, J., Kinyanjui, E. M., & Ong'ele, S. (2018). Scaling up successfully: Lessons from Kenya's Tusome national literacy program. Journal of Educational Change, 19(3), 293–321. https://doi.org/10.1007/s10833-018-9325-4.

140. Piper, B., Simmons Zuilkowski, S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. World Development, 106, 324–336. https://doi.org/10.1016/j.worlddev.2018.01.018.

141. Ponce, J. & Bedi, A. S. (2010). The impact of a cash transfer program on cognitive achievement: The Bono de Desarrollo Humano of Ecuador. Economics of Education Review, 29(1), 116–125. https://doi.org/10.1016/j.econedurev.2009.07.005.

142. Powell, C. A., Walker, S. P., Chang, S. M., & Grantham-McGregor, S. M. (1998). Nutrition and education: a randomized trial of the effects of breakfast in rural primary school children. The American Journal of Clinical Nutrition, 68(4), 873–879. https://doi.org/10.1093/ajcn/68.4.873.

143. Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia. American Economic Journal: Applied Economics, 6(2), 105–126. https://doi.org/10.1257/app.6.2.105.

144. Pugatch, T. & Schroeder, E. (2018). Teacher pay and student performance: evidence from the Gambian hardship allowance. Journal of Development Effectiveness, 10(2), 249–276. https://doi.org/10.1080/19439342.2018.1452778.

145. Reinikka, R. & Svensson, J. (2011). The power of information in public services: Evidence from education in Uganda. Journal of Public Economics, 95(7–8), 956–966. https://doi.org/10.1016/j.jpubeco.2011.02.006.

146. Rodríguez, C., Sánchez, F., & Armenta, A. (2010). Do Interventions at School Level Improve Educational Outcomes? Evidence from a Rural Program in Colombia. World Development, 38(3), 415–428. https://doi.org/10.1016/j.worlddev.2009.10.002.

147. Romero, M., Sandefur, J., & Sandholtz, W. A. (2020). Outsourcing Education: Experimental Evidence from Liberia. American Economic Review, 110(2), 364–400. https://doi.org/10.1257/aer.20181478.

148. Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., Grau, V., Lagos, F., López, X., López, V., Rodriguez, P., & Salinas, M. (2003). Beyond Nintendo: design and assessment of educational video games for first and second grade students. Computers & Education, 40(1), 71–94. https://doi.org/10.1016/s0360-1315(02)00099-4.

149. Sabates, R., Rose, P., Alcott, B., & Delprato, M. (2021). Assessing cost-effectiveness with equity of a programme targeting marginalised girls in secondary schools in Tanzania. Journal of Development Effectiveness, 13(1), 28–46. https://doi.org/10.1080/19439342.2020.1844782.

150. Sailors, M., Hoffman, J. V., Pearson, P. D., Beretvas, S. N., & Matthee, B. (2010). The Effects of First- and Second-Language Instruction in Rural South African Schools. Bilingual Research Journal, 33(1), 21–41. https://doi.org/10.1080/15235881003733241.

151. Sampa, F. K., Ojanen, E., Westerholm, J., Ketonen, R., & Lyytinen, H. (2018). Literacy programs efficacy for developing children's early reading skills in familiar language in Zambia. Journal of Psychology in Africa, 28(2), 128–135. https://doi.org/10.1080/14330237.2018.1435050.

152. Santibañez, L., Abreu-Lastra, R., & O'Donoghue, J. L. (2014). School based management effects: Resources or governance change? Evidence from Mexico. Economics of Education Review, 39, 97–109. https://doi.org/10.1016/j.econedurev.2013.11.008.

153. Simeon, D. T., Grantham-McGregor, S. M., Callender, J. E., & Wong, M. S. (1995). Treatment of Trichuris trichiura Infections Improves Growth, Spelling Scores and School Attendance in some Children. The Journal of Nutrition, 125(7), 1875–1883. https://doi.org/10.1093/jn/125.7.1875.

154. Singh, A., Romero, M., & Muralidharan, K. (2024). COVID-19 Learning loss and recovery. Journal of Human Resources, 0723-13025R2. https://doi.org/10.3368/jhr.0723-13025r2.

155. Solon, F. S., Sarol, J. N., Bernardo, A. B. I., Solon, J. A. A., Mehansho, H., Sanchez-Fermin, L. E., Wambangco, L. S., & Juhlin, K. D. (2003). Effect of a Multiple-Micronutrient-Fortified Fruit Powder Beverage on the Nutrition Status, Physical Fitness, and Cognitive Performance of Schoolchildren in the Philippines. Food and Nutrition Bulletin, 24(4_suppl_1), S129-S140. https://doi.org/10.1177/15648265030244s110.

156. Spears, D. & Lamba, S. (2016). Effects of Early-Life Exposure to Sanitation on Childhood Cognitive Skills: Evidence from India's Total Sanitation Campaign. Journal of Human Resources, 51(2), 298–327. https://doi.org/10.3368/jhr.51.2.0712-5051r1.

157. Stampini, M., Martinez-Cordova, S., Insfran, S., & Harris, D. (2018). Do Conditional Cash Transfers Lead to Better Secondary Schools? Evidence from Jamaica's PATH. World Development, 101, 104–118. https://doi.org/10.1016/j.worlddev.2017.08.015.

158. Stern, J. M. B., Jukes, M. C. H., Cilliers, J., Fleisch, B., Taylor, S., & Mohohlwane, N. (2026). Persistence and Emergence of Literacy Skills: Long-Term Impacts of an Effective Early Grade Reading Intervention in South Africa. Journal of Research on Educational Effectiveness, 1–22. https://doi.org/10.1080/19345747.2024.2417288.

159. Tan, J., Lane, J., & Lassibille, G. (1999). Student Outcomes in Philippine Elementary Schools: An Evaluation of Four Experiments. The World Bank Economic Review, 13(3), 493–508. https://doi.org/10.1093/wber/13.3.493.

160. Trani, J., Zhu, Y., Bechara, S., Bakhshi, P., Kaplan, I., Babulal, G., Zha, W., Rawab, H., Brown, D., & Raghavan, R. (2025). The impact of a participatory intervention to improve learning outcomes and reduce school-based discrimination and community stigma in primary rural schools of Afghanistan: A cluster control randomized trial. International Journal of Educational Development, 118, 103409. https://doi.org/10.1016/j.ijedudev.2025.103409.

161. Urquiola, M. (2006). Identifying Class Size Effects in Developing Countries: Evidence from Rural Bolivia. Review of Economics and Statistics, 88(1), 171–177. https://doi.org/10.1162/rest.2006.88.1.171.

162. Wang, L. C., Vlassopoulos, M., Islam, A., & Hassan, H. (2024). Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh. Journal of Political Economy Microeconomics, 2(3), 562–601. https://doi.org/10.1086/730456.

163. Wang, C., Xiao, A., & Zhou, Y. (2024). Teamwork and Human Capital Development. Journal of Human Resources, 59(5), 1425–1457. https://doi.org/10.3368/jhr.0121-11400R2.

164. Watkins, W. E., Cruz, J., & Pollitt, E. (1996). The effects of deworming on indicators of school performance in Guatemala. Transactions of the Royal Society of Tropical Medicine and Hygiene, 90(2), 156–161. https://doi.org/10.1016/s0035-9203(96)90121-2.

165. Weldeegzie, S. (2023). The persistent effect of conflict on educational outcomes: Evidence from Ethiopia. International Journal of Educational Development, 103, 102884. https://doi.org/10.1016/j.ijedudev.2023.102884.

166. Wolf, S., Aber, J. L., Behrman, J. R., & Tsinigo, E. (2019). Experimental Impacts of the "Quality Preschool for Ghana" Interventions on Teacher Professional Well-being, Classroom Quality, and Children's School Readiness. Journal of Research on Educational Effectiveness, 12(1), 10–37. https://doi.org/10.1080/19345747.2018.1517199.

167. Wong, H. L., Luo, R., Zhang, L., & Rozelle, S. (2013). The impact of vouchers on preschool attendance and elementary school readiness: A randomized controlled trial in rural China. Economics of Education Review, 35, 53–65. https://doi.org/10.1016/j.econedurev.2013.03.004.

168. Yamauchi, F. (2014). An alternative estimate of school-based management impacts on students' achievements: evidence from the Philippines. Journal of Development Effectiveness, 6(2), 97–110. https://doi.org/10.1080/19439342.2014.906485.

169. Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M. C., Weiland, C., Gomez, C. J., Moreno, L., Rolla, A., D'Sa, N., & Arbour, M. C. (2015). Experimental impacts of a teacher professional development program in Chile on preschool classroom quality and child outcomes. Developmental Psychology, 51(3), 309–322. https://doi.org/10.1037/a0038785.

170. Zhang, L., Lai, F., Pang, X., Yi, H., & Rozelle, S. (2013). The impact of teacher training on teacher and student outcomes: evidence from a randomised experiment in Beijing migrant schools. Journal of Development Effectiveness, 5(3), 339–358. https://doi.org/10.1080/19439342.2013.807862.

171. Zhang, X., Wang, L., & Ye, X. (2025). Combating Learning Poverty: Experimental Evidence on Improving Instruction Through Teacher Training. Economics of Education Review, 109, 102729. https://doi.org/10.1016/j.econedurev.2025.102729