CENTER FOR GLOBAL DEVELOPMENT

Phonics and Foreign Aid: Can America Teach the World to Read?

Justin Sandefur, Thomaz Alvares de Azevedo, Xiaomin Ju, and and Thi Le

Abstract

Over two decades and dozens of countries, the United States Agency for International Development has refined a package of support for early-grade reading, often referred to as "structured pedagogy," which includes textbooks, teacher training, coaching, and lesson plans. Programs are implemented by American companies in public schools and evaluated using harmonized learning metrics, yielding a portfolio of 12 experimental and 15 difference-in-difference evaluations. Results vary widely, but on average programs increase oral reading fluency by approximately 3 words from a base of 13 correct words per minute in early primary. The average program costs about $200 per pupil, roughly equivalent to doubling school spending. Larger programs cost much less per pupil, but yield (insignificantly) smaller impacts. Newer programs yield somewhat bigger impacts, consistent with the idea that program evaluation can improve the quality of aid through iterative learning.

KEYWORDS
education, early-grade reading assessments, impact evaluation, foreign aid, USAID

JEL CODES
F35, I25, O15

WORKING PAPER 668 · NOVEMBER 2023

Phonics and Foreign Aid: Can America Teach the World to Read?

Justin Sandefur
Center for Global Development

Thomaz Alvares de Azevedo
MSI, A Tetra Tech Company

Xiaomin Ju
Center for Global Development

Thi Le
Center for Global Development

We're grateful to Alicia Menendez, Amy Mulcahy-Dunn, Matthew Jukes, Nancy Murray, Ben Piper, and Elena Walls for helpful comments and for answering numerous questions. This work was supported by grants from the Bill & Melinda Gates Foundation and Open Philanthropy. All errors and opinions are ours alone, and do not represent the views of the authors' employers, the Center for Global Development's board or its funders.

Justin Sandefur, Thomaz Alvares de Azevedo, Xiaomin Ju, and and Thi Le. 2023. "Phonics and Foreign Aid: Can America Teach the World to Read?" CGD Working Paper 668. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/phonics-and-foreign-aid-can-america-teach-world-read

CENTER FOR GLOBAL DEVELOPMENT
2055 L Street, NW Fifth Floor
Washington, DC 20036
202.416.4000

1 Abbey Gardens
Great College Street
London
SW1P 3SE

www.cgdev.org

Center for Global Development. 2023.

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

Contents

1 Introduction 4

2 Program background and evaluation designs 8

2.1 Interventions 9

2.2 Evaluation designs 9

2.3 Outcomes 11

2.4 Timing and samples 12

3 Treatment effect estimates: How much do USAID programs improve reading levels? 13

3.1 Specification 13

3.2 Program-by-program results 15

3.3 Diagnostic tests 16

3.4 Meta-analytic methods 18

3.5 Meta-analytic results: moderate, positive, significant average effects with high variance across programs 19

4 External validity: Do projects without rigorous evaluations have similar performance? 21

4.1 Adjusting first-differences in reading outcomes for changes in the timing of tests 22

4.2 Comparing first-differences in the treatment group between studies with and without a control group 23

4.3 Back-of-the-envelope estimates of the magnitude of the bias 24

5 Cost effectiveness: Comparing learning per dollar over time, by program size, and to regular government spending 26

5.1 Economies of scale 26

5.2 Comparison to other education interventions and business-as-usual education spending 27

6 Can fidelity of implementation explain results across schools (or projects)? 29

6.1 READ Liberia: big learning gains, moderate changes in classroom practice . 31

6.2 Kenya PRIMR: big learning gains, big changes in classroom practice 31

6.3 Uganda SHRP: modest learning gains, big changes in classroom practice . . 32

6.4 Kyrgyzstan QRP: modest learning gains, moderate changes in classroom practice 33

7 Is mother-tongue instruction effective? 34

7.1 Comparing language outcomes within mother-tongue programs 35

7.2 Comparing programs with and without mother-tongue instruction 35

7.3 Comparing pupils who do and don't speak curricular languages at home . . . 36

8 Conclusion 37

List of Figures

1 Map of early-grade reading evaluations by methodology 49

2 Impacts on oral ready fluency 50

3 Average oral reading fluency by grade and month, pooling all studies 51

4 Changes in oral reading fluency in the treatment group (only) 52

5 Average unit cost versus scale 53

6 The impact of mother-tongue instruction 54

7 Impacts on oral ready fluency, including implied effects for programs with no control group 64

8 Impacts on oral ready fluency – expressed in standard deviation effect sizes 65

9 Impacts on oral ready fluency – expressed in 'equivalent years of schooling' . 66

10 Impacts on correct letter sounds per minute 67

11 Test for sorting of pupils in response to treatment: treatment effects on pupil age 68

12 Test for publication bias 69

List of Tables

1 Project meta-data . . . 55

2 Intervention variants . . . 56

3 Summary statistics & balance . . . 57

4 Treatment effects on oral reading fluency . . . 58

5 Meta-regressions . . . 59

6 Intermediate outcomes . . . 60

7 Associations between intermediate outcomes and learning gains . . . 61

8 Meta-analysis results for average treatment effects . . . 62

9 Incorporating before-and-after studies into overall treatment effect estimate 62

10 Timing of data collection and school calendar by program . . . 63

1 Introduction

Decades of experimental research have established that kids learn to read faster when literacy instruction emphasizes phonemic awareness (breaking words into distinct sounds), synthetic phonics (connecting letters to sounds and blending them together), and guided oral reading (National Reading Panel (US) and National Institute of Child Health and Human Development (US), 2000). Since the mid-2000s, America's flagship foreign aid program in basic education has focused on encouraging teachers in developing countries to adopt these evidence-based approaches to teaching reading. The United States Agency for International Development (USAID) has refined a standard package of support, sometimes referred to as "structured pedagogy", whose elements include teacher in-service training, pre-written lesson plans or teaching guides, follow-up coaching and mentoring for teachers, instructional materials for students, and tools for student assessment. A recent high-level panel of experts classified the approach as one of the three most cost-effective interventions to improve learning outcomes in the developing world (Banerjee et al., 2023).

The package involves minimal transfer of money or goods to developing countries, apart from the books. USAID contracts American companies to deliver the program in each country, targeting anywhere from one thousand to several million primary grade learners per program, at a cost of between USD $2 million and $165 million per program.

How cost effective is this form of foreign aid at improving learning outcomes in developing countries? This paper attempts to answer that question by re-analyzing microdata from 8 experimental and 31 non-experimental USAID evaluations of early-grade reading programs across 29 countries (portions of our meta-analysis also incorporate published results from 4 additional randomized trials, for which microdata is not publicly available). We build on a recent effort within USAID to assess the performance of the agency's full global portfolio early-grade reading programs from 2011 to 2021 (Mulcahy-Dunn and Alvares de

Azevedo, 2023). USAID's approach to procuring and evaluating programs provides three distinct analytical advantages for making generalizable inferences about cost-effectiveness: (i) mandatory evaluations of all education programs, (ii) harmonized outcome metrics across programs, and (iii) transparent top-line budget numbers for programs targeting well-defined populations.

First, from 2011 onward USAID policy has required evaluations of all programs, rather than just innovative or successful ones.[1] This allows us to talk about the impact of a meaningful universe of interventions: USAID's early-grade reading programs. Despite concerns in the education literature that treatment effects attenuate when programs are taken to scale (Bold et al., 2018; Kerwin and Thornton, 2021), the evidence here is ambiguous: we observe a negative but statistically insignificant relationship between program scale and impact. There is also a significant tendency for 'better' (and richer) programs to receive more rigorous evaluations. Specifically, evaluations with a control group are more likely to be found in programs with higher spending per pupil, and faster learning gains within treatment schools. This implies that systematic reviews or meta-analyses that filter studies by evaluation design, e.g., Graham and Kelly (2019) in the case of early-grade reading programs, may inadvertently exaggerate average impacts.

Analyzing a well-defined universe of programs also circumvents publication bias, and the tendency for larger effects to be more prominently cited. This ends up being quantitatively important. For instance, a recent high-level panel of experts classified USAID's "structured pedagogy" approach as one of the three most cost-effective interventions to improve learning outcomes in the developing world (Banerjee et al., 2023). The panel's report highlights two USAID programs, in Kenya (Piper et al., 2014) and Liberia (Menendez et al., 2023) respec-

[1] In USAID's 2011 evaluation policy, evaluations were mandated for awards larger than the average for the operating unit. From 2020 onwards, the threshold was fixed at \$20 million in total value. The requirement that education programs, specifically, measure learning outcomes was enshrined in law by the 2017 READ Act.

tively, which we show are the largest impacts in the USAID portfolio. But the expert panel omits similar USAID programs that produced null effects, including both quasi-experimental evaluations in Indonesia and Malawi (RTI, 2017; Tilson et al., 2013) and randomized evaluations in Uganda and Kyrgyzstan. [2]

The second unique advantage presented by USAID's global evaluation portfolio is the use of harmonized outcome metrics. This allows us to partially circumvent a common problem in the empirical education literature, where it is widely recognized that effect sizes on learning do not permit meaningful comparisons across studies, particularly in developing countries where there is a dearth of standardized tests and item banks to draw from (Bertling et al., 2023). In contrast, USAID requires all evaluations to measure impacts on oral reading fluency using a common tool known as an early-grade reading assessment (EGRA), with well-documented protocols (Gove and Wetterberg, 2011). A key remaining threat to comparability is differences in word length across languages, which we address in part by reporting effects on an additional harmonized metric, correct letter sounds/names per minute (Abadzi, 2012).

To preview the results on this harmonized scale, impacts are modest on average but highly variable across programs. Random effects meta-analytic estimates of the average impact across experimental and (non-experimental) difference-in-differences evaluations represent a learning gain of about 3 correct words per minute. Estimates range from a drop in oral reading fluency of -6.1 correct words per minute in the DRC to a gain of 14.6 in Liberia. Observable program characteristics, including scale, unit cost, etc. do little to explain this variation. In contrast to Angrist and Meager's (2023) findings for TaRL, we find little evidence that implementation quality can explain variation in results for the handful of

[2] (Evans and Popova, 2016) highlights a related problem, showing that systematic reviews of what works in education for development reach widely different conclusions, due in part to differences in inclusion criteria, but also to subjective ex post classification of studies into intervention buckets. USAID's ex ante commitment to evaluate all early-grade reading programs offers a potential solution to both these challenges.

projects where such data is available. We stress, however, that implementation metrics here are sparse and do not capture implementation quality in much detail, so any claims for or against the importance of implementation fidelity in explain program variation remain fairly speculative.

Rigorous evaluations (including randomized trials) are non-randomly placed. Comparing progress in the treatment group across studies, we find that improvements in learning outcomes within programs schools are significantly faster in studies with a control group (4.28 CWPM improvement) than in studies without (2.68 CWPM improvement). We calculate that this phenomenon, which is potentially widespread in the program evaluation literature but usually impossible to observe, may bias overall treatment effects upward by roughly half, albeit from a small base in this case.

A third analytical advantage of USAID's evaluation portfolio is that reliance on federal procurement systems to award grants and contracts renders cost figures unambiguous and transparent. This is a consequence of the way USAID operates. Because all programs consist of an arm's length transaction between USAID and an awardee (almost always an American company) for a specific early-grade reading activity with a defined beneficiary population, we can directly observe the funder's marginal cost for the activity in question. Arguably, this provides a more credible, transparent measure of program costs than is usually possible when organizations attempt to estimate their own costs for an individual activity.

The average proram costs around $200 per pupil per year, which is roughly equivalent to doubling per pupil education expenditure in these school systems. Average spending per pupil across the whole portoflio, however, is much lower, at around $34. This discrepancy is due to the fact that larger programs, reaching up to several million pupils, report sharply lower unit costs. While this may suggest possible economies of scale, as already noted, larger programs also have (insignificantly) smaller treatment effects.

The following section describes USAID's early grade reading portfolio in more detail, before discussing our estimation approach, meta-analytic methods and presenting the core results for the RCTs and difference-in-differences studies in Section 3. We then broaden our scope in Section 4 to include an additional 16 evaluations with no control group, finding slower learning gains in these programs. Cost data is reported in Section 5, while Section 6 reports data on implementation fidelity from a subset of programs where it is available. Section 7 explores the role of one specific program component, mother-tongue instruction, and Section 8 concludes.

2 Program background and evaluation designs

USAID's reading programs are what is known as "tied" aid, routed through companies in the donor country. No funds flow directly to governments, companies, NGOs, or schools in the developing world.[3] While the United States contributes on the order of $10 billion per annum to multilateral aid organizations (including those specializing in education), a roughly equal amount is awarded to American intermediaries to implement programs – including all of the early-grade reading programs covered here.[4] Specifically, fourteen of the 49 programs were administered wholly or partially by Research Triangle International, a non-profit based in North Carolina; seven by Creative Associates, a for-profit company based in Maryland; and seven by Education Development Center, a non-profit based in Massachusetts.

Geographically, USAID's early-grade reading work focuses heavily on sub-Saharan Africa (see map in Figure 1). Twenty-nine of the 49 total evaluations from 2011 to 2021 fell in this region, followed by nine in Asia, six in Latin America, four in the Middle-East and North Africa, and one in Eastern Europe.

[3] In a handful of cases, the USAID has subsequently made grants to governments to institutionalize successful USAID reading programs. Those efforts are not covered in the evaluations studied here.

[4] Authors' calculations based on foreignassistance.gov.

2.1 Interventions

The majority of the programs in USAID's early-grade reading portfolio targeted grades 1 through 3, with some extending to earlier and/or later grades as well. Table 1 provides a full listing.

The core of most programs is teacher training, usually on a phonics-based curriculum, and distribution of teaching and learning materials. Graham and Kelly (2019) code a sample of projects (including several USAID programs included here, as well as other non-USAID projects) on whether or not they include each of five intervention categories: (i) training teachers on evidence-based curricula; (ii) providing instructional guidelines; (iii) following up with coaching and mentoring; (iv) providing instructional materials; and (v) providing tools and training for student assessment. In Table 2 we add a sixth to this list, mother-tongue instruction, and code the remainder of the USAID programs. Almost all programs include the first, third, and fourth components.

The content of programs has evolved over time, e.g., with a stronger emphasis on teacher coaching, particularly in the wake of the release of 2013 USAID Strategy Implementation Guidance which directed the design of future early grade reading programs. Earlier programs were more structured (i.e., more prescriptive for teachers) and less adapted to local contexts (Mulcahy-Dunn and Alvares de Azevedo, 2023). Later programs place greater emphasis on cooperation with ministries of education. Notably, later programs have also benefited from the creation of learning materials designed and piloted in earlier programs in the same country as well as availability of gender-disaggregated data on learning outcomes.

2.2 Evaluation designs

In total, we are aware of 49 evaluations of USAID early-grade reading programs. Data files are unavailable in some cases and published results are insufficient to compare to our analysis, leaving us with 43 studies included in our meta-analysis. In 39 cases we re-analyze microdata, relying on published RCT results for the remaining four.

Notably, USAID requires that implementation and evaluation of a given program is awarded to separate organizations.

In terms of research design, the evaluations can be divided into three groups. First, 13 programs underwent randomized control trials. All of these, except for one in the Kyrgyz Republic and one in Tajikistan, took place in Africa (two in Uganda, and one each in Kenya, Ethiopia, Liberia, Mozambique, and Nigeria) or Latin America (two in Peru, and one each in Guatemala and Nicaragua). Randomization was done at the school level, in nine of thirteen cases, and at a higher 'cluster' or 'zone' level in three cases: Kenya PRIMR, Uganda SHRP, and Uganda LARA; Nicaragua EpC randomly assigned different units—children or educational communities—depending on the size of the educational community Baseline, pre-treatment data is available in seven of thirteen cases. In most cases, however, pupils and often schools are re-sampled at endline, so there is no longitudinal panel of students (Kenya PRIMR and the LAC Reads studies are exceptions here, as well as sub-samples of the Kyrgyz Republic QRP and Tajikistan QRP studies).

Second, an additional 15 programs were evaluated using a difference-in-differences design, collecting outcomes from both treatment and (non-randomized) control schools before and after treatment. In two cases, we observe statistically significant baseline imbalance (see Table 3): in Zambia, treatment schools report lower reading scores by 2 correct words per minute at baseline, and in Indonesia they report 5 correct words per minute higher scores. None of these diff-in-diff evaluations include more than one pre-treatment round of data collection, so it is impossible to test assumptions about parallel pre-trends. And as in the case of the randomized evaluations, these non-experimental diff-in-diff datasets are repeated cross-sections of pupils rather than pupil panels, though sometimes include a panel of schools.

Finally, the remaining 19 programs are evaluated on the basis of a simple time-series comparison of treated schools before and after treatment, without any control schools. Once again, pupils (and often schools) are re-sampled for each round, thus data sets are not pupil-level panels, which is a virtue in this case: by returning to children in a given grade, sampled from the sample population of schools and classes, these evaluations seek to measure whether learning outcomes are improving over time conditional on grade.

2.3 Outcomes

We focus on oral reading fluency (ORF) as the primary outcome, and specifically, the number of correct words per minute (CWPM) a child is able to read. ORF measures students' ability to read aloud, and USAID funded the development of detailed protocols for how to measure it (Gove and Wetterberg, 2011). While average ORF scores are often treated as the primary outcome in program reports, evaluations also frequently highlight the percentage of pupils with a score of zero CWPM, or who clear thresholds like 10 or 40 CWPM.

When all children are tested in multiple languages, we report results for both (e.g., English and Kiswahili in Kenya). When children are tested in one of multiple languages (as in Kyrgyzstan where pupils were tested in either Kyrgyz or Russian, or Uganda where the USAID implementing partner delivered programs in either Luganda, Lunyankore/Rukiga, or Runyoro-Rutooro), We report pooled results controlling for the language of testing.

One challenge is that some languages use fewer, longer words to say the same thing. For instance, Abadzi (2012) reports data from Matthew Jukes showing that a similar reading passage in a first-grade textbook in Kenya contains roughly 60 English words compared to just 40 Kiswahili words, due to the agglutinative nature of Kiswahili. This poses obvious problems for comparing correct words per minute across languages. We attempt to circumvent this problem in two ways. First, to check robustness of cross-language comparisons, we report all results in standard deviation learning gains. Standardization should remove baseline differences in words per minute across languages, but introduces other challenges for comparability (different samples have different standard deviations, so results are arguably not in comparable units).

Second, following a suggestion in Abadzi (2012), we also report treatment effects on a secondary outcome which is not affected by agglutination: correct letters per minute rather than words. This outcome presents a lower bar for pupils, which has the additional advantage of producing less censoring or bottom coding (in some samples over 90 percent of pupils read zero correct words per minute). To maximize coverage, we pool results from studies that use two slightly different measures: correct letters per minute (i.e., naming the letter) and correct letter sounds per minute (articulating the sound they make phonetically). Studies generally report either one or the other, not both.

2.4 Timing and samples

All evaluations include repeated cross-sectional samples of pupils from a given grade or grades (80 percent in grades 1, 2, and/or 3, and the remainder in grades 4, 5, and 6). [5] We restrict attention to impacts in the last available round of outcome data. Many evaluations include baseline data, one or more rounds of midline data, and a final round of endline data. We report results for the endline round, or latest midline round for which microdata is available, and use only baseline, pre-treatment (never midline, post-treatment) measures as controls.

Where control groups exist, data collection is always synchronized between treatment and control groups, which is a necessary condition for the validity of the treatment effects we estimate. However, various factors including elections, teacher strikes, etc. sometimes

[5] We ignore the longitudinal sub-samples of pupils tracked over time in some evaluations as, to the best of our knowledge, only one study tracked the full set of both treatment and control schools, and the relevant longitudinal identifiers are not available.

force the timing of data collection to change between survey rounds. This does not bias our (diff-in-diff) treatment effects estimates, but does complicate the interpretation of simple single-difference estimates from the before-and-after studies as discussed in detail in Section 4.

3 Treatment effect estimates: How much do USAID programs improve reading levels?

3.1 Specification

The econometric specification necessarily varies with the evaluation design.

1. **Diff-in-diff with school FE.** Whenever feasible,we report results from a difference-in-differences specification, regressing oral reading fluency for pupil i in school j at time t on the interaction between treatment status and the post-treatment (endline) data round, as well as controls for school fixed effects, time period, and a vector of student characteristics, $\mathbf{X}_{ijt}$, including gender, language of testing, grade, and age.[6]

$$ORF_{ijt} = \beta Treated_j \times Endline_t + \eta_j + u_t + \gamma \mathbf{X}_{ijt} + \varepsilon_{ijt} \tag{1}$$

   Here the coefficient of interest is β, the coefficient on the interaction between treatment status and a dummy for the post-treatment round of data collection. This specification applies equally to randomized or non-randomized evaluations.

[6] Socio-economic status (SES) is likely to explain a considerable amount of pupil-level variation in reading outcomes. To date we have not harmonized SES variables across evaluations, but hope to make progress on this in future drafts.

2. **Diff-in-diff without school FE.** In some cases, both pre- and post-treatment data is available, but either (i) the sample of schools changes across survey rounds, or (ii) school identifiers are not provided in the data. In these cases, equation (1) is amended, replacing the school fixed effects, η_j, with an indicator for schools (eventually) assigned to treatment, $\eta Treated_j$.

3. **Cross-sectional comparison of treatment and control schools.** When baseline data is not available, but samples include both treatment and control schools, we estimate a simple cross-sectional regression:

$$ORF_{ij} = \beta Treated_j + \gamma \mathbf{X}_{ij} + \varepsilon_{ij} \tag{2}$$

   Notably, this case arises only in the context of randomized control trials.

4. **Before-and-after with school FE.** A substantial number of evaluations failed to collect any outcome data on non-treated schools whatsoever (in some cases because all schools in the country were theoretically treated). In these cases, we report estimates from the following school-level panel regression:

$$ORF_{ijt} = \beta Endline_t + \eta_j + \gamma \mathbf{X}_{ijt} + \varepsilon_{ijt} \tag{3}$$

   Once again, β is the coefficient of interest, measuring the improvement in reading scores over time within schools, controlling for observable changes in the composition of students.

5. **Before-and-after without school FE.** Finally, as with the diff-in-diff specification, in some cases it is not possible to identify the same schools across survey rounds, and we estimate a version of equation (3) without school fixed effects. The β coefficient has the same interpretation, but may be subject to greater sampling error.

For randomized studies, standard errors are clustered at the level of randomization. This is generally the school level, but sometimes a larger aggregation such as a cluster, zone, or district. For non-randomized studies, standard errors are clustered at the level of the primary sampling unit. (In principle, one might wish to cluster at the level of non-random treatment assignment, but this is rarely reported in practice.)

3.2 Program-by-program results

The raw treatment effects for each individual study are reported in Table 4 and Figure 2. Among the experimental studies, impacts on oral reading fluency range from a maximum of 8.9 correct words per minute for English reading in Kenya's PRIMR program (p-value = 0.01) to 0.2 correct words per minute for English reading in Uganda's SHRP program (p-value = 0.92). The Kyrgyzstan QRP program yields a fairly precise null result, and the OPEC program in the Democratic Republic of Congo is also insignificant. Otherwise, all studies yield significant, positive impacts on oral reading fluency, albeit of varying magnitudes.

Among the non-experimental difference-in-differences studies, the largest impact is a slight outlier, represented by a treatment effect of 11.5 correct words per minute in Ghanaian local languages for Ghana's PFE Learning program (p-value < 0.01). The same program yielded an effect of 4.3 correct words per minute in English (p-value < 0.01). The smallest effect among the diff-in-diff studies was -6.1 correct words per minute in French in the PAQUED program in the Democratice Republic of Congo (p-value = 0.03).

Note that the goal of the analysis here is not to replicate or validate the estimates originally reported in USAID reports. In many cases the specification employed here is different (e.g., reporting wherever possible a difference-in-differences specification, including school fixed effects, clustering standard errors at the level of randomization where applicable, and restricting our focus to pre-treatment and the final post-treatment data collection round, as well as pure control groups and full treatment arms).

Finally turning to the results for correct *letters* per minute, Figure 10 shows the forest plot of treatment effects from individual studies. For the randomized trials, impacts range from -1.7 correct letters per minute (p-value $= 0.19$) in the Kyrgyzstan QRP program to 16.5 CLPM (p-value < 0.01) improvement in English reading in the Kenya PRIMR program. For the non-randomized difference-in-differences studies, impacts range from -6.3 CLPM (p-value $= 0.04$) in the Malawi MTPDS program to 21 CLPM (p-value < 0.01) improvement in local language reading in the Ghana PFE Learning program.

Apart from their independent interest, one value of these results for correct letters per minute is that they provide a rough check on whether the ranking of program impacts is driven by word length in specific languages. Impacts on correct words per minute may be smaller in agglutinative languages with longer words, but impacts on correct letters per minute should not be. Across these two outcome metrics, we find that the ranking of the outcomes from 21 experimental and difference-in-differences programs for which we have both measures shows a Spearman rank correlation of approximately 0.7. This broad alignment of results across the two outcomes, although not perfect, gives us some confidence that program comparisons are not driven by language differences.

3.3 Diagnostic tests

One generic concern with the validity of these results is possible changes in the composition of pupils in response to treatment. In most cases, pupils are not tracked over time. All of the analysis here relies on repeated cross-sections of pupils, usually drawn from a partially overlapping set of schools pre- and post-treatment, but sometimes from an entirely new set of schools. This sampling strategy will induce bias if pupils enroll, drop out, or change schools in response to treatment, or—because sampling is done at the grade level—if schools change their repetition and promotion policies in response to treatment.

To test for this behavior, we estimate a variant of the main treatment effects specification (equations (1) - (3)), but using pupil age as the dependent variable. If, for instance, treatment leads schools to hold kids back a grade until they master certain reading skills, this should show up as a positive treatment effect on pupil age. Alternatively, if more affluent, more academically prepared, younger pupils flock to treatment schools, this may manifest as a negative treatment effect on age. The results, shown in Figure 11, are generally reassuring. For the 26 evaluations which record pupil age, we find significant evidence of changes in pupil composition in response to treatment in just one. (Of course, this does not rule out other changes in composition unrelated to age.)[7]

A closely related concern, which may not be picked up by our test for age differences, is endogenous school *attendance* in response to treatment. Because samples are mostly drawn from pupils in attendance on the day of testing, treatment effects would be biased downward (for instance), if the program encourages more marginal students to attend, driving down scores in treatment schools. We have little ability to test this hypothesis, but flag it as a caveat in the interpretation here.

A second potential concern in reviewing any portfolio of evaluations such as this, is the possibility of publication bias. In principle, USAID's commitment to independent evaluation for both learning and accountability purposes should mitigate this risk: the agency produces evaluations to ensure implementing partners do good work, rather than to market the agency's successes. This transparency appears to be borne out in the data. Figure 12

[7] The exception is the DRC Accelere program, which shows a negative treatment effect on pupil age. All pupils were sampled in grade 5. Results are sensitive to how one deals with 50 cases of missing age, all in the baseline control group. Controlling for or omitting these missing values produces the effect shown: treatment pupils are somewhat younger at endline. This phenomenon is potentially linked to the simultaneous roll out of the government's free education policy, but the 0.5 year decline in pupil age specific to treatment schools here remains unexplained.

shows a funnel plot of all the estimated treatment effects, with standard errors on the vertical access and effect size on the horizontal axis. One tell-tale sign of publication bias is an asymmetric distribution of coefficients, with 'missing' results closer to zero. The top panel shows no evidence of this asymmetry. Another possible sign of publication bias is clustering of coefficients just beyond p-value significance thresholds. The bottom panel shows no clear pattern of this form of bias either.

While there is little evidence of changes in pupil composition or of publication bias, the evaluations have some design flaws which are worth noting in interpreting these results. In addition to the lack of longitudinal pupil panels, the public use data files often lack school identifiers, making it impossible to control for school fixed effects or (in some cases) even to cluster standard errors at the school level. Furthermore, there is very little discussion of non-compliance at the school level in any of the evaluation reports. We interpret all estimates as ITT effects, incorporating non-compliance with the program curriculum by teachers. For the experimental evaluations, details on the process of randomization are fairly limited. Deviations from randomization, if any, are not documented, though USAID monitoring and evaluation staff report that such cases should be extremely rare.

3.4 Meta-analytic methods

Given the variation in context and program design across studies, we favor a random effects model over the fixed effects approach. The random effects model allows for variance across studies in the true effect, such that

$$\hat{\beta}_j = \beta + u_j + \varepsilon_j, \text{ where } \varepsilon_j\ N(0, \sigma^2), u\ N(0, \tau^2) \tag{4}$$

The weights used to estimate the overall sample average become $w_j = 1/(\hat{\sigma}_j^2 + \hat{\tau}_j^2)$. We report empirical Bayes estimates of $\hat{\tau}^2$ (see Viechtbauer et al. (2015) for details) in the forest plots.

Either the fixed effects or random effects model can be extended to allow for meta-analytic controls in a regression framework, i.e., controls for context- or study-specific variables which moderate the effect observed in study j.

To test the hypothesis that true effects vary across studies, we report a test of residual homogeneity, where null is $\tau^2 = 0$. Under the null, the test statistic

$$Q = \sum_{j=1}^{K} w_j(\hat{\beta}_j - \hat{\beta})^2$$

follows a χ^2 distribution with $K - 1$ degrees of freedom.

Several programs report results for more than one language, e.g., English and Kiswahili in Kenya, or English and one of several possible local languages in Ghana. To handle the correlated effect sizes, we use the robust variance estimation (RVE) method developed by Hedges et al. (2010) with small sample adjustments suggested by Tipton (2015). We use the RVE results as our main results.

3.5 Meta-analytic results: moderate, positive, significant average effects with high variance across programs

On average, USAID reading programs produced significant, positive learning gains of about 3 correct words per minute. Estimated impacts are strikingly similar across experimental and non-experimental evaluations, but vary widely across countries. That variation doesn't appear to be correlated with program size, cost, or time of implementation.

Beginning with average effects, across all studies, pooling both experimental and non-experimental results in a single random effects model, the mean effect size was 2.9 correct words per minute (p-value < 0.01). Breaking this down by evaluation design, the average effect across experimental studies was 3.06 CWPM (p-value < 0.01) as show in Figure 2, the average effect for diff-in-diff designs was 2.75 CWPM (p-value $= 0.01$).

As noted above, the estimated effects span a wide range, from -6.1 to 11.5 CWPM. Unsurprisingly, the data reject the null hypothesis of homogeneity across studies (see again Figure 2). This is true across all studies, and within both experimental studies and non-experimental diff-in-diff studies. In each case the null is rejected with a p-value < 0.01.

To test the robustness of the overall average effect sizes, Table 8 reports results using a fixed effects model as well as our benchmark random effects model. Average impacts remain right around 3 correct words per minute. We also report estimates weighting studies purely by the number of beneficiaries (which may differ dramatically from the sample size). That change reduces the average effect across RCTs and diff-in-diff studies to about 2 correct words per minute.

Turning to correct letter names and letter sounds per minute (pooled in the bottom panel of Table 8), we calculate a mean effect size of 4.52 correct letters per minute (p-value < 0.01). This varies little by evaluation type, and is once again smaller if we weight studies by the number of beneficiaries rather than the inverse variance of estimates. Recall that a key motivation of including letters as well as words as an outcome metric was to ensure results across projects are not driven by differences in word length across languages. Somewhat reassuringly, the best (and worst) projects on one outcome tend to be the best (or worst) on the other. The Spearman rank correlation of effect sizes between outcomes is 0.69 across programs.

Basic program characteristics provide few clues as to why some produce so much bigger impacts than others. Meta-analytic regressions reported in Table 5 show no significant association between effect sizes and program size (measured by the log of total beneficiary pupils), log cost per pupil, indicators of whether the program was experimental or quasi-experimental (as opposed to uncontrolled), and when the program occurred.

One possible interpretation of these results is that contextual factors matter more than program characteristics in explaining variation in results. Implementers report, e.g., differential levels of government enthusiasm and "buy in" as a key factor explaining program success. In any case, USAID implements a fairly standardized early-grade reading model across countries, but impacts vary quite widely.

4 External validity: Do projects without rigorous evaluations have similar performance?

Are the results reported above a good guide to the overall impact of USAID early-grade reading programs globally? The experimental and (to a slightly lesser degree) non-experimental difference-in-differences evaluations provide internally valid estimates of treatment effects from a given program. But these evaluations account for just over half the evaluations in our sample, and the placement of randomized evaluations (and more broadly, evaluations with any kind of control group) is likely to be non-random.

Allcott (2015) documents the phenomenon of "site-selection bias" in the case of randomized experiments in energy conservation. In that context, programs themselves are targeted at populations with greater need (higher energy use) or greater demand (more environmentalist areas), both of which produce larger treatment effects in early trials compared to later replications.

Here we examine a slightly different phenomenon. Our hypothesis is that programs that saw larger learning gains in the treatment group and spent more money per pupil were more likely to be evaluated using a randomized trial or difference-in-differences design.

4.1 Adjusting first-differences in reading outcomes for changes in the timing of tests

We explore whether the findings from rigorous evaluations are likely to be externally valid for other programs by comparing studies with a control group to studies without, testing for differences in learning gains over time in the treatment group. Focusing on first differences in this way requires careful attention to the timing of data collection. As noted above, survey timing is frequently dictated by events outside the control of USAID or the research teams, with activities planned for the start of the school year pushed to the middle of the year, and so on.

In most of the before-and-after studies with no control group, baseline and endline data collection occur roughly in the same month of the school year (+/- 1 month in 14 of 16 cases), usually one or two years apart. Recall, these studies are generally returning to a repeated cross-section of pupils in the same grade(s), so estimates measure changes in reading for, say, second-graders in April of one year and of a new crop of second-graders in roughly April of the following year.

There are cases, however, where baseline and endline data collection occur at different times of the year. In particular, in the studies with a control group, it is common to observe baseline data collection toward the beginning of the year, followed by endline data collection toward the end of the same or subsequent school years. And in practice, intermediate cases exist, complicating interpretation of descriptive statistics about first-differences. (See Table 10 in the appendix for a full list of survey timing by program.)

To allow for meaningful interpretation of first-differences, we adjust all learning outcomes for the month of data collection. This amounts to adding or subtracting a month effect to each survey round. (Note that we do this adjustment only for these comparisons of first-differences across evaluation methods, and not for the main treatment effect estimates above, where we rely on raw outcome data and require only that surveys are synchronized between the treatment and control arm.)

We base this month-of-survey adjustment on an auxiliary, pooled regression (using microdata from all studies) of oral reading fluency on the number of months since the start of the school year, controlling for grade dummies, program fixed effects, and treatment effects (i.e., the interaction of being in the treatment group and being observed at endline). We also interact months since the start of the school year with grade, to allow the trajectory of learning over time to change as pupils progress through primary school. Letting i denote pupils, j schools, k studies, and t time periods, we estimate:

$$ORF_{ijkt} = \sum_{g=1}^{6} I[Grade_{ijkt} = g] \sum_{m=1}^{12} I[Month_{ijkt} = m] + \beta Treated_j \times Endline_t + \phi_k + \varepsilon_{ijkt} \quad (5)$$

The estimated grade and month dummies are depicted in Figure 3, and used to adjust scores to simulate the result of doing all testing in the first month of each school year. Results imply that students gain, on average across all samples, about 1.3 CWPM per month in grades 1 to 3, and lose about 10 CWPM over the summer (see Figure 3). In upper grades, the sample is smaller and results are less precisely estimated, but learning gains appear to flatten off and summer learning loss disappears. The important caveat here is that these results are based on repeated cross-sections rather than longitudinal panels (and so subject to some composition effects if pupils join or drop out).

4.2 Comparing first-differences in the treatment group between studies with and without a control group

First-differences in oral reading fluency are shown in Figure 4, which measures the change in correct words per minute from baseline to endline for children in the same grade.

Looking at the before-and-after studies with no control group, the largest estimate is an increase of 7.5 correct words per minute in English under Kenya's Tusome program (p-value = 0.01), which was the successor to the PRIMR program which produced large gains in the experimental phase.[8] At the other extreme, Arabic reading levels fell by -1.2 correct words per minute in Lebanon during the period of the QITABI program (p-value = 0.01).

The average change in literacy for before-and-after studies was 2.68 correct words per minute (see the first column of Table 9). Results were much stronger for programs subject to a difference-in-differences evaluation, where reading increased by 4.01 correct words per minute in the treatment group, and yet stronger again in programs subject to experimental evaluation, where reading increased by 6.48 correct words per minute.

In short, programs with faster learning gains were more likely to receive rigorous evaluations. This may be partially explained by per pupil spending levels, which are considerably lower in programs evaluated with only before-and-after data. [9] In some cases, experimental and non-experimental evaluations happen in sequence as the program is scaled up, in which case our results imply smaller effects in the later, larger, non-randomized programs. Though note that we do not find a significant association between treatment effects and program scale overall.

4.3 Back-of-the-envelope estimates of the magnitude of the bias

To provide a sense of how much this may bias estimates of average program impacts, we explore the implications of assuming that the *change over time* in performance of control

[8] Freudenberger and Davis (2017) find higher learning levels at the midline of the program, which is not evaluated here. Both midline and endline data collection happened 2-3 months later in the school year than baseline data collection, complicating the interpretation of results. We adjust for this timing difference at endline as described in the text above.

[9] As a caveat, note that this difference in program spending by evaluation type is somewhat fragile to the choice of weights. Using the weights applied throughout the paper for the meta-analysis, programs with rigorous evaluations spend much more per pupil than those without. But if we weight programs by the number of beneficiaries, this ceases to be true.

schools does not differ systematically by evaluation type. This is untestable, as we do not observe control schools in before-and-after studies. The assumption would be violated if—and to reiterate, we assume this is not the case, and USAID staff have assured us this is not the case—more rigorous evaluations were assigned to places already experiencing broad-based improvement in education quality independent of the program.

We estimate an average improvement in control schools of 1.9 correct words per minute across all experimental and difference-in-differences studies (see the second column of Table 9). For illustrative purposes, we subtract this average first-difference among control schools (from experimental and difference-and-differences studies) from the first-difference in learning outcomes in treatment schools (in before-and-after studies) to simulate treatment effects for the latter group. This yields an average treatment effect across before-and-after schools of 1.09 correct words per minute (p-value = 0.25).

Pooling these simulated treatment effects from before-and-after studies with the estimated treatment effects from experimental and difference-in-differences studies yields an overall average impact of 2.22 correct words per minute, compared to 2.9 correct words per minute from the experimental and difference-in-differences studies alone (see Figure 7 in the appendix). We interpret the former as an estimate of the average impact of USAID early-grade reading programs, and the gap as an estimate of the upward bias in treatment effects induced by the non-random placement of rigorous evaluations.

5 Cost effectiveness: Comparing learning per dollar over time, by program size, and to regular government spending

The fact that USAID contracts out all of its early-grade reading programs facilitates cost comparisons. Conveniently, these programs are largely stand-alone awards from USAID to an implementing partner or consortium of partners. Implementers are contracted and funded by USAID to carry out activities, and generally provided limited supplemental budgets from outside sources, so total the total USAID award is a reasonable approximation of total project cost. We divide this figure by the total number of pupils in treated classrooms to calculate average per pupil costs.

Our cost metrics omit some complementary, in-kind contributions from host-country governments which are not recorded and would be difficult to cost. This rarely includes significant material resource transfers. Rather, even when programs are implemented by American intermediaries, American intermediaries work with government staff to implement these activities.

5.1 Economies of scale

The median USAID early-grade reading program reaches half a million pupils (total) over the course of three school years, at a total cost of USD $38 million. In pupil terms, the smallest program was an experiment in Nigeria with just six thousand students, and the largest was the nationwide scale-up of the Tusome program in Kenya, reaching nearly 8 million pupils. In dollar terms, Macedonia had the smallest program ($1.7 million), and Pakistan the largest ($164.7 million).

It appears that many of the costs associated with these programs are fixed, creating strong economies of scale. As shown in Figure 5, a simple regression fit suggests average per-pupil costs fall by 0.7 log points for every log point increase in the number of pupils treated. The extremes in terms of unit cost were Egypt with a per pupil expenditure of $3,167 for 12,000 students, and Malawi with a per pupil cost of just $6 spread over 3.2 million students (though as we will see, the latter produced somewhat disappointing learning outcomes).

5.2 Comparison to other education interventions and business-as-usual education spending

Is 3 words per minute a lot? There are various ways to think about this subjective characterization. Expressed in effect sizes (i.e., multiples of a pupil standard deviation in the control group and/or baseline data), learning gains are about 0.3 standard deviations. (See Figure 8 in the appendix for project specific treatment effects in standard deviations.) But the use of standard deviations in oral reading fluency in many contexts covered here is somewhat dubious. In the extreme case, the standard deviation of oral reading fluency for the control group at baseline in the DRC OPEQ program was zero. Zero pupils in the sample could read a single word. This is not an entirely unique problem: in seven different programs, over 90 percent of pupils could read zero words at baseline.

An alternative approach, sometimes employed in the education policy literature, is to express results in terms of the typical increase in learning associated with an additional year of schooling (Evans and Yuan, 2019). We calculate the benchmark learning pace by regressing oral reading fluency on pupil grade in the cross-section for each study, acknowledging the limitations of this approach.[10] Focusing on the experimental and difference-in-differences

[10] First, policies such as grade repetition for lagging students will bias this measure of normal learning progress. Second, small treatment effects in an absolute sense will appear large in systems where normal progress is especially slow.

evaluations, average impacts in our sample (aggregating with the random effects meta-analytic model) are about .43‘equivalent years of schooling’ (p-value = 0.02). See Figure 9 in the appendix for full results.

Bringing costs into this calculation, we continue to use ‘business as usual’ in public schools as a benchmark. On the benefit side, the equivalent years of schooling metric compares treatment effects to normal learning progress.[11] Thus on the cost side, we compare the cost of the program per pupil to government spending per pupil per annum, taken from the World Bank’s World Development Indicators.

$$\begin{aligned}\text{Relative cost effectiveness} &= \frac{\text{Equivalent Years of Schooling}}{\text{Ratio of program cost to gov't spending}} \\ &= \tfrac{\text{Treatment effect}}{\text{Normal learning pace per year}} \Big/ \tfrac{\text{Program cost per pupil}}{\text{Gov't spending per pupil}}\end{aligned}$$

Values greater than 1 imply program spending is more cost-effective than existing government expenditure at improving reading.

Consider what this calculation yields for the RCTs and difference-in-differences evaluations only. Treatment effects are about 3 correct words per minute, and a student normally gains about 9 words per minute in oral reading fluency per year of schooling in early primary. So effects are equivalent to the learning normally acquired in 0.3 additional years of schooling. On the cost side, average program cost is around $200 per pupil, similar to average government expenditure per pupil. Combining these metrics separately for each project and then averaging using the same weights used in our meta-analysis of treatment effects, results are slightly better than that aggregate would imply. The average result for the relative cost

[11] Recall from Section 2 that we avoid double counting studies that measure impacts on two languages. An objection might be that teaching two languages has cost implications, and so double counting is appropriate. One advantage of the relative cost effectiveness formula used here is that this issue of double counting becomes arithmetically irrelevant. Whether we credit a USAID program with two sets or learning gains (in, say, English and a local language) or the average of those two gains makes no difference to the result, so long as we use the same attribution rule when constructing our business-as-usual benchmark in control schools.

effectiveness measure from the equation above is around 1.1 in the RCTs and diff-in-diff studies. In other words, USAID programs generate about 110 percent as much learning per dollar spent as regular government expenditure on education.

This comparison is crude in multiple respects. It treats USAID programs and regular government expenditure as competing priorities, whereas USAID spending is actually layered on top of government spending. Our calculation also takes the association between learning and years of schooling as causal. Perhaps less importantly, we ignore the opportunity cost of children's time, which might favor more learning per days of schooling over more days, and also ignore the non-reading gains (and non-reading expenditures) in both normal school systems and USAID reading programs, including math learning and socio-emotional development. Nevertheless, the results provide a rough order of magnitude for the cost effectiveness of the USAID early-grade reading portfolio, which appears roughly similar to regular government spending.

6 Can fidelity of implementation explain results across schools (or projects)?

An obvious potential explanation for the wide variation in performance across (erstwhile similar) USAID reading programs is that there are likely variations in implementation quality. Angrist and Meager (2023) show that while intent-to-treat estimates of the impact of "teaching at the right level" interventions across five programs vary widely, once implementation is effectively controlled for in treatment-on-the-treated estimates, they cannot reject homogeneous effects across programs. In our setting, the point of the training, coaching, and other interventions at the core of the USAID model is that they will change what goes on in the classroom. That cannot be taken for granted. In this section, we explore the extent to which that actually happened. We focus on experimental programs to avoid the risk of trying to explain illusory effects (causal mediation is hard enough as it is), and deliberately choose both successful and unsuccessful cases.

The results here are somewhat inconclusive. While USAID imposes harmonization of learning outcomes across programs, there is no such harmonization in terms of what intermediate outcomes are measured. Some programs included systematic classroom observation using an established rubric and independent enumerators, while others add a few questions to the pupil survey asking kids what they do in class.

We present results on implementation and potential mechanisms in two steps. First, Table 6 reports treatment effects on intermediate outcomes, including variables related to (a) reading pedagogy and classroom practice, (b) learning assessment and monitoring of student progress, and (c) availability of books and learning materials. The basic hypothesis is that USAID programs, if working as intended, should move the needle on each of these three intermediate outcomes, and lack of impact in some programs may be due to a simple failure to change these basic practices/conditions.

Second, Table 7 tests whether these same intermediate indicators are indeed associated with learning outcomes. The table also reports a version of the benchmark treatment-effects regression from equation (1) in which we include intermediate outcomes. This common approach to mediation analysis, sometimes referred to as “the product method” (VanderWeele, 2016) requires strong assumptions to identify causal mediators. Imai et al. (2010) refer to these assumptions jointly as “sequential ignorability”, consisting of an assumption about conditional independence of the treatment (delivered by randomization) and of the mediator conditional on treatment (which is not guaranteed in our context). Any interpretation of the results in Table 7 should be read with these assumptions in mind.

6.1 READ Liberia: big learning gains, moderate changes in classroom practice

The experimental evaluation of Liberia's latest READ program produced the largest gain of any study in the USAID portfolio. Surprisingly however, (Menendez et al., 2023) find little evidence that classroom practice improved much at all. Based on a combination of qualitative observation and mediation regressions, they attribute learning gains to the presence of teaching and learning materials in the classroom.

Estimates in Table 6 show that the READ Liberia program led to a 20 percentage point (pp) increase in the share of pupils who say they have books at school that they can take home to read, a 13 pp increase in the share who say the teacher made them practice reading out loud in class, an 18 pp increase in the share who say their teacher assigns reading to do at home, and a 16 pp increase in the share who say their teacher ever makes them re-tell a story during class. There is no significant impact, however, on the share who report their teacher making them practice silent reading in class.

6.2 Kenya PRIMR: big learning gains, big changes in classroom practice

The Kenya PRIMR study is another experimental evaluation with fairly big positive impacts on reading fluency (Piper et al., 2014), though an additional ICT component appeared to have no marginal impact (Piper et al., 2016). PRIMR has garnered considerable international attention in part due to the national scale up of a similar intervention, rebranded as TUSOME – see positive learning results for both studies above. Piper et al. (2018) describe lessons from that scale-up process.

Systematic classroom observation data show a number of significant changes in teacher practice under PRIMR as shown in Table 6. Students spend more time reading both aloud and silently, more time on letters and sounds, and teachers spend more time quizzing them on reading comprehension. Observers also report more use of the textbook during class.

Nevertheless, including these factors in the main treatment effects regression has little explanatory power: the size and significance of the coefficient on the PRIMR treatment dummy is robust to controling for these intermediate outcome measures, as shown in Table 7 — indeed, if anything, the estimate coefficient increases in magnitude after controlling for these hypothetical channels.

6.3 Uganda SHRP: modest learning gains, big changes in classroom practice

In Uganda's SHRP program, estimated treatment effects on reading fluency were small (in the case of local languages) or null (in the case of English).

USAID commissioned a qualitative follow-up study in the wake of these disappointing RCT findings, focusing on four factors (Brunette et al., 2019). That ex post review found high compliance with the teacher training component (i.e., teachers in treatment schools had actually attended training), teachers gave largely positive feedback about the training, but complained that applying it in real classes was challenging. Compliance with teacher support supervsion was also reasonably good: all teachers reported receiving at least some supervision. Feedback on the usefulness of this supervision was mixed. Absenteeism among both pupils and teachers was put forward as a major impediment. Finally, the component of the program encouraging community involvement in the school was generally deemed a failure by teachers.

These conclusions roughly match what we see in Table 6, which shows SHRP led to large and statistically significant changes in a number of relevant practices: teachers are more likely to guide students to read from printed materials, teachers rely more on textbooks, learners are more likely to have textbooks, and teachers are more likely to have records of learning assessments. Strikingly, possession and use of textbooks or printed materials by pupils in the classroom is perfectly collinear with treatment: it was observed in no control classroom and every treatment classroom. (Hence, it is impossible to estimate the role of these factors in a mediation-style regression in Table 7.)

6.4 Kyrgyzstan QRP: modest learning gains, moderate changes in classroom practice

The Kyrgyzstan QRP program is a second example of a large-scale, experimental evaluation with null effects on reading fluency. In this case, the available intermediate indicators offer mixed evidence about how much changed inside classrooms, although the range of variables is limited.

The pupil survey reports just four variables related to the intermediate outcomes of interest. As shown in Table 6, the QRP program had no impact on the share of pupils with a reading textbook or the share reporting reading homework, which were both almost 100 percent in the control group. QRP did have a significant positive impact on the share of pupils who reported the teacher checking their reading work (28 pp increase) and the share reporting a reading activity outside of class (54 pp increase). Perplexingly though, these intermediate outcomes show a significant negative association with learning gains in Table 7.

Stepping back, a key take-away from this section is that it is difficult to reach any firm conclusions about the role of implementation fidelity in explaining variation in program outcomes in USAID's early-grade reading portfolio due to the lack of systematic data on implementation. Where indicators exist, they may not adequately capture the quality of implementation. For now, there is little ground to claim that implementation differences can explain differences in program impacts. Given the paucity of data though, it is important to remember that absence of evidence (that implementation fidelity is a key factor here) is not evidence of absence (i.e., it would be premature to conclude that it is not).

7 Is mother-tongue instruction effective?

There is a long-standing debate in the literature questioning whether pupils' long-run literacy is best served by bilingual instruction or immersion in a new language (Slavin and Cheung, 2005). This literature has found little systematic difference between the two approaches, but has been dominated by research on Spanish-language instruction in the United States, and those lessons may or may not be universally applicable elsewhere.

For the purposes of this section, we treat the USAID programs in Uganda and Ghana as mother-tongue programs, in that they added new languages of instruction over-and-above what is already used in government schools, to accommodate the home language of learners. We compare these to programs such as those in Liberia or Mozambique where a large share of learners do not speak the language of instruction at home, but mother-tongue instruction was not part of the program.[12]

None of the evaluations provide a side-by-side comparison of mother-tongue instruction and non-mother-tongue instruction. However, looking both across the portfolio of evaluations, and within some specific programs, we can shed some light on the effectiveness of mother-tongue instruction as part of USAID's early-grade reading programs.

[12] Notably, we place USAID's Kenya programs in the latter group: they include Kiswahili instruction, as do regular public school classes, but do not include other languages commonly spoken by students at home.

7.1 Comparing language outcomes within mother-tongue programs

Within the same program, treatment effects are bigger in mother tongue than other languages. Three experimental or difference-in-differences evaluations report impacts on oral reading fluency in both mother-tongue and an official language: Ghana PFE Learning, Uganda LARA, and Uganda SHRP. Brunette et al. (2019) show the effectiveness of the Uganda SHRP program in raising mother-tongue reading levels for several Uganda languages. In both Uganda programs and in Ghana, we find the impacts are bigger for mother-tongue outcomes (though not significantly so in the case of Uganda SHRP), and bigger by about 3 correct words per minute on average, as shown in Figure 6a.

This is perhaps unsurprising. Both the treatment and control group are being tested in a language which the control group has never been taught to read. Nevertheless, these results demonstrate that it was possible to make significant gains in mother-tongue reading when and where it was introduced.

7.2 Comparing programs with and without mother-tongue instruction

We now return to the hypothesis that introducing mother-tongue instruction can accelerate literacy acquisition in other curricular languages (e.g., English or French). An important caveat here is that all of our results are fairly short-term. While we fail to see strong evidence for this hypothesis during the duration of the program, it might be more reasonable to expect the benefits of mother-tongue instruction in early-primary to manifest themselves only in later grades, which we cannot test here.

Across programs, those which introduced mother-tongue instruction produced statistically indistinguishable impacts on reading levels in other languages (see Figure 6b. Here we limit our focus to programs where a significant proportion of the population does not speak the language of instruction at home. This includes Ghana, Uganda, and Zambia where USAID program schools introduced mother tongue instruction, as well as the Democratic Republic of Congo (where the USAID program was in French), Indonesia (where the program was in Indonesian), Kenya (English and Kiswahili), Liberia (English), Mozambique (Portuguese), and Yemen (Modern Standard Arabic). Clearly reading improvements in mother tongue may be a policy goal in their own right, and comparisons of treatment effects across a handful of programs provides only minimal, suggestive evidence (at best) of the marginal effect of specific program elements. Nevertheless, we observe that among these countries with a large number of pupils who don't speak the language of instruction at home, the programs that introduced mother-tongue instruction saw average reading gains of 3.68 correct words per minute (p-value = .02) in the original language of instruction (i.e., not in their mother tongue), while those which did not employ mother-tongue instruction saw gains of 3.28 correct words per minute (p-value = 0). This difference is statistically insignificant, i.e., we find no evidence here that introducing mother-tongue instruction helps or hinders literacy acquisition in the original curricular language.[13]

7.3 Comparing pupils who do and don't speak curricular languages at home

Next we take a closer look at two programs which underwent RCTs and did *not* introduce mother-tongue instruction, despite the fact that a large share of pupils do not speak the language of instruction at home. In Kenya, PRIMR included instruction in both English and Kiswahili, but 40 percent of pupils reported speaking a different language at home

[13] Note that the estimates here are based on all pupils in the sample, including those whose mother tongue does or does not match the language of instruction. In the next section we distinguish these two groups.

(predominantly Kikuyu or Luo), while in Liberia instruction under the READ program was exclusively in English, which 73 percent of pupils said they did not speak at home.

In both cases, reading gains were indistinguishable between students who spoke the language of instruction at home and those who did not. In Kenya, students who reported not speaking either English or Kiswahili at home experienced slightly smaller learning gains (2.5 fewer additional correct words per minute, p-value = 0.20), while in Liberia these students experienced slightly larger learning gains (0.7 additional correct words per minute, p-value = 0.64).

8 Conclusion

Building on ten years of impact evaluations of USAID reading programs around the world, this paper attempts to harmonize and aggregate impacts on oral reading fluency. This evaluation portfolio provides a test case of a common form of foreign aid: provision of technical assistance by foreign firms and NGOs to improve service delivery in the developing world. Because USAID evaluations are intended for both accountability as well as learning, they span a large, coordinated set of programs accounting for billions of dollars in aid across dozens of countries. Evaluations are not limited to pilots but extend to programs 'at scale' with millions of beneficiaries and attendant implementation challenges.

Averaging across all evaluations with a plausible control group in a meta-analytic framework, we find effects of approximately 3 correct words per minute on oral reading fluency among children in early primary school. Those estimated effects are roughly the same whether considering randomized trials or non-exerpimental difference-in-difference evaluations. There are signs that impacts may be smaller in programs without such rigorous evaluations. Both the RCT and DiD evaluations show larger learning improvements in the treatment group than found in the 16 additional programs reporting only before-and-after data with no control group.

Is 3 words per minute a lot? There are various ways to think about this subjective characterization. Expressed in effect sizes (i.e., multiples of a pupil standard deviation in the control group and/or baseline data), learning gains are about 0.3 standard deviations. But the use of standard deviations in oral reading fluency in many contexts covered here is somewhat dubious. In the extreme case, the standard deviation of oral reading fluency for the control group at baseline in the DRC OPEQ program was zero. Zero pupils in the sample could read a single word. This is not an entirely unique problem: in seven different programs, over 90 percent of pupils could read zero words at baseline. Expressing results in terms of the learning associated with an additional year of schooling in a regular public school, average impacts are about 0.3 'equivalent years of schooling'.

Results vary widely across programs and countries, spanning the range from significant negative impacts on oral reading fluency to double-digit gains in correct words per minute. This variation significantly exceeds the level one would expect due to mere sampling variation across seemingly similar programs (i.e., statistical tests for homogeneity of effects reject at the one-percent level).

Observable program characteristics provide limited clues as to why impacts vary so much. Impacts decline with project scale, but this relationship is not statistically significant. Measures of program implementation – e.g. the extent to which teachers actually use the phonics techniques they are trained in – are largely inconclusive. Existing metrics of implementation fidelity have little explanatory power within or across programs. But these measures are limited to a handful of projects, and are often quite coarse.

Reading gains tend to be larger in the mother tongue. But it is not the case that programs focusing on mother-tongue instruction produce larger gains in official languages over the relatively short time-frame covered in the evaluations.

We interpret the variance in impacts as suggesting that context matters for program performance, and stress the need to tailor both programs and expectations to contextual variation. We note, however, that alternative explanations cannot be fully ruled out, including variation in efficiency across contractors and unobserved details of program design beyond the limited factors we've explored here. Additional data collection on both the quality of program implementation and contextual factors that explain performance might increase the value of future evaluations.

The programs are expensive by local standards. Using the same weighting approach used to calculate average impacts, the average program cost is over $200, or roughly similar to routine government expenditure per primary pupil. Stated differently, this implies USAID early-grade reading programs are equivalent to doubling current expenditure in exchange for a 30 percent increase in reading gains on average. Larger programs cost much less per pupil though, bringing average spending down to just $34 per pupil when weighted by project size. Cost effectiveness thus hinges on the ability of projects to scale up while maintaining effect sizes.

Finally, it is worth noting that all the evidence reported here reflects learning outcomes during or very shortly after the completion of USAID programs. There is no indication as to whether these impacts will persist in either of two senses: (i) whether children who experience learning gains in lower-primary will go on to higher educational attainment and achievement, and (ii) whether teachers and schools whose pedagogical practices improved under the program will continue to produce higher learning levels over coming years. This lack of evidence on sustainability is particularly noteworthy given USAID's heavy reliance on American implementing partners. Future work could examine the long-run impacts of USAID's more recent efforts to embed programs more meaningfully within recipient-country school systems.

References

Helen Abadzi. Developing cross language metrics for reading fluency measurement: Some issues and options. Global Partnership for Education (GPE) Working Paper Series on Learning No. 6, 2012.

J. Lawrence Aber, Leighann Starkey, Carly Tubbs, Catalina Torrente, Brian Johnston, Sharon Wolf, Anjuli Shivshanker, and Jeannie Annan. Opportunities for equitable access to quality basic education (opeq). final report on the impact of the opeq intervention in the democratic republic of congo, 2015. URL https://www.rescue.org/sites/default/files/document/642/ed-opportunitiesforequitableaccesstoqualitybasiceducation.pdf. Authoring organization: New York University (NYU), International Rescue Committee (IRC). Sponsoring organization: United States Agency for International Development. Document type: Final Report.

Elizabeth Adelman, Adrienne Lucas, and Genta Menkulasi. Usaid kenya (education for marginalized children in kenya) (november 2006 – december 2014) fy 2006 - 2014 – emack project endline evaluation report (october 2012 – december 2014), 2014. Authoring organization: External Consultants to Aga Khan Foundation. Sponsoring organization: United States Agency for International Development. Document type: Endline Evaluation Report.

AIR. Usaid reading for ethiopia's achievement developed monitoring and evaluation(read m&e). early grade reading assessment (egra) 2018 endline report, 2019. URL https://2017-2020.usaid.gov/sites/default/files/documents/1865/Ethiopia-Early-Grade-Reading-Assessment-2018.pdf. Sponsoring organization: United States Agency for International Development. Document type: Endline Report.

AIR and Save the Children. Usaid quality reading project kyrgyz republic: Final egra and impact report 2013–2017, 2017a. Sponsoring organization: United States Agency for International Development. Document type: Final Impact Report.

AIR and Save the Children. Usaid quality reading project republic of tajikistan. final egra and impact report 2013–2017, 2017b. Sponsoring organization: United States Agency for International Development. Document type: Final Impact Report.

Hunt Allcott. Site selection bias in program evaluation. The Quarterly Journal of Economics, 130(3):1117–1165, 2015.

Noam Angrist and Rachael Meager. Implementation matters: Generalizing treatment effects in education (june 21, 2023). Available at SSRN https://ssrn.com/abstract=4487496, 2023.

Emilie Bagby, Catalina Torrente, Steve Glazerman, Nancy Murray, and Ivonne Padilla. Impact evaluation of espacios para crecer (epc), an afterschool program in nicaragua. final report, 2019. URL https://pdf.usaid.gov/pdf_docs/PA00XDR4.pdf. Authoring organization: Mathematica. Sponsoring organization: United States Agency for International Development (USAID). Document type: Evaluation, Final Report.

Abhijit Banerjee, Tahir Andrabi, Rukmini Banerji, Susan Dynarski, Rachel Glennerster, Sally Grantham-Mcgregor, Karthik Muralidharan, Benjamin Piper, Jaime Saavedra Chanduvi, Hirokazu Yoshikawa, Sara Ruto, and Sylvia Schmelkes. 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? World Bank http://documents.worldbank.org/curated/en/099420106132331608/IDU0977f73d7022b1047770980c0c5a14598eef8, 2023.

Wajd Baraheem. Yemen early grade reading – yegr endline survey. final report, 2013. Authoring organization: Prodigy Systems. Sponsoring organization: . Document type: Endline Survey Final Report.

Masha Bertling, Abhijeet Singh, and Karthik Muralidharan. Psychometric quality of measures of learning outcomes in low- and middle-income countries. Center for Global Development, Working Paper 638 https://www.cgdev.org/publication/psychometric-quality-measures-learning-outcomes-low-and-middle-income-countries, 2023.

Tessa Bold, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. Experimental evidence on scaling up education reforms in kenya. Journal of Public Economics, 168:1–20, 2018.

Tracy Brunette, Benjamin Piper, Rachel Jordan, Simon King, and Rehemah Nabacwa. 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, 2019.

Juanita (Jennie) Campos, Janet K. Orr, Benedicta C. Agusiobo, and Hadiza Shettima. Northern education initiative plus (nei+): Mid-term evaluation, 2017. URL https://pdf.usaid.gov/pdf_docs/PA00N894.pdf. Authoring organization: DevTech Systems, Inc.. Sponsoring organization: United States Agency for International Development. Document type: Mid-term Evaluation.

Larissa Campuzano, Camila Fernandez, Julieta Lugo-Gil, Steve Glazerman, Nancy Murray, and Ivonne Padilla. Evaluation of amazonia lee reading intervention in peru. final report, 2018. URL https://www.edu-links.org/sites/default/files/media/file/amazonia%20lee%20final%20report.pdf. Authoring organization: Mathematica. Sponsoring organization: United States Agency for International Development (USAID). Document type: Evaluation, Final Report.

Creative Associates. Afghan children read. early grade reading assessment in dari language. conducted in the province of herat. revised midline report, 2019. URL https://earlygradereadingbarometer.org/pdf/Afghanistan%20Midline%20Revised%20EGRA%20reprt%20in%20Dari%20language,%20Afghan%20Children%20Read%20(ACR)%20Project.pdf. Document type: Midline Report.

EdIntersect and Chemonics. Lecture pour tous/all children reading senegal early grade reading assessment/ snapshot of school management effectiveness 2021 report, 2021. URL https://pdf.usaid.gov/pdf_docs/PA00Z6BF.pdf. Sponsoring organization: United States Agency for International Development. Document type: Assessment.

Education Development Center. Literacy, language and learning intitiative (l3) national fluency and mathematics assessment of rwandan schools endline report, 2017. URL https://l3.edc.org/documents/EDC-L3-Endline-Evaluation.pdf. Sponsoring organization: United States Agency for International Development. Document type: Endline Report.

Evaluating Systems. Ghana early grade reading program impact evaluation — 2019 endline report, 2019. URL https://pdf.usaid.gov/pdf_docs/PA00XHTT.pdf. Sponsoring organization: United States Agency for International Development. Document type: Endline Report.

David Evans and Fei Yuan. Equivalent years of schooling: A metric to communicate learning gains in concrete terms. World Bank Policy Research Working Paper, (8752), 2019.

David K Evans and Anna Popova. What really works to improve learning in developing countries? an analysis of divergent findings in systematic reviews. The World Bank Research Observer, 31(2):242–270, 2016.

Zachariah Falconer-Stout, Rebecca Frischkorn, and Lynne Miller Franco. Time to learn endline evaluation report, 2017. URL https://pdf.usaid.gov/pdf_docs/PA00MK22.pdf. Authoring organization: EnCompass LLC.. Sponsoring organization: United States Agency for International Development (USAID), the USAID/Zambia Read to Succeed Project. Document type: Endline Evaluation Report.

Elizabeth Freudenberger and Jeff Davis. Tusome external evaluation–midline report, 2017. URL https://pdf.usaid.gov/pdf_docs/PA00MS6J.pdf. Authoring organization: Management Systems International. Sponsoring organization: United States Agency for International Development. Document type: Evaluation Midline Report.

Zewdu Gebrekidan. Second early grade reading assessment of level ii learners in selected abecs of amhara, oromia, somali and tigray regions, 2014. Authoring organization: PACT Ethiopia. Sponsoring organization: United States Agency for International Development. Document type: Assessment.

Ghana Education Service, National Education Assessment Unit, RTI, and Education Assessment and Research Centre. Ghana 2015 early grade reading assessment and early grade mathematics assessment: Report of findings, 2016. URL https://ierc-publicfiles.s3.amazonaws.com/public/resources/Ghana%202015%

20EGRA-EGMA_22Nov2016_FINAL.pdf. Sponsoring organization: United States Agency for International Development. Document type: Assessment, Report of Findings.

Amber K Gove and Anna Wetterberg. The early grade reading assessment: Applications and interventions to improve basic literacy. RTI Press, 2011.

Jimmy Graham and Sean Kelly. How effective are early grade reading interventions? a review of the evidence. Educational Research Review, 27:155–175, 2019.

Larry V Hedges, Elizabeth Tipton, and Matthew C Johnson. Robust variance estimation in meta-regression with dependent effect size estimates. Research synthesis methods, 1(1): 39–65, 2010.

IBTCI. Monitoring, evaluation, and coordination contract: Accelere! activity1 reading impact evaluation report, 2020. Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation Report.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. Psychological methods, 15(4):309, 2010.

IQPEP. Third iqpep-early grade reading assessment (final report), 2014. Document type: Final Report.

Jennifer Iriondo-Perez, Maureen Kelly, Jennae Bulat, Aarnout Brombacher, and Timothy Slade. Paqued: Drc. projet d'amélioration de la qualité de l'education (paqued): 2014 endline report of early grade reading assessment (egra) and early grade mathematics assessment (egma), 2014. URL https://pdf.usaid.gov/pdf_docs/pa00mgj3.pdf. Authoring organization: RTI International. Sponsoring organization: United States Agency for International Development, Education Development Center. Document type: Endline Report.

Erika Keaveney, Carlos Fierros, Alexander Rigaux, and Alicia Menendez. Tusome external evaluation endline report, 2020. URL https://www.norc.org/PDFs/Tusome%20Endline%20Performance%20Evaluation/PA00XVBP.pdf. Authoring organization: NORC at the University of Chicago. Sponsoring organization: United States Agency for International Development. Document type: Evaluation Endline Report.

Jason T Kerwin and Rebecca L Thornton. 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, 2021.

Simon King, Medina Korda, Lee Nordstrum, and Susan Edwards. Liberia teacher training program: Endline assessment of the impact of early grade reading and mathematics interventions, 2015. URL https://shared.rti.org/content/liberia-teacher-training-program-endline-assessment-impact-early-grade-reading-and#modal-21-512. Authoring organization: RTI International. Sponsoring organization: United States Agency for International Development. Document type: Endline Assessment of Impact.

Jeremy Koch, Sarah Fuller, Elizabeth Freudenberger, Dana Kelly, Jeff Davis, Idalia Rodriguez Morales, and Sean Kelly. 2017 early grade reading assessment balochistan, 2018. URL https://pdf.usaid.gov/pdf_docs/PA00TNK7.pdf. Authoring organization: Management Systems International. Sponsoring organization: United States Agency for International Development. Document type: Assessment.

Julieta Lugo-Gil, Nancy Murray, Camila Fernandez, Steven Glazerman, and Larissa Campuzano. Evaluation of leer juntos, aprender juntos early grade reading intervention in guatemala. final report, 2019a. URL https://pdf.usaid.gov/pdf_docs/PA00XJKG.pdf. Authoring organization: Mathematica. Sponsoring organization: United States Agency for International Development (USAID). Document type: Evaluation, Final Report.

Julieta Lugo-Gil, Nancy Murray, Steven Glazerman, Camila Fernandez, and Ivonne Padilla. Evaluation of leer juntos, aprender juntos early-grade reading intervention in peru. final report, 2019b. URL https://pdf.usaid.gov/pdf_docs/PA00XJKM.pdf. Authoring organization: Mathematica. Sponsoring organization: United States Agency for International Development (USAID). Document type: Evaluation, Final Report.

Alicia Menendez and Gregory Haugan. Impact evaluation of the early grade reading program (egrp) in nepal. midline evaluation report, 2018. URL https://pdf.usaid.gov/pdf_docs/PA00Z4B1.pdf. Authoring organization: NORC at the University of Chicago. Sponsoring organization: United States Agency for International Development. Document type: Midline Evaluation Report.

Alicia Menendez, Ursula Hoadley, and Anna Soloyeva. Read liberia impact evaluation endline report, 2021. URL https://pdf.usaid.gov/pdf_docs/PA00Z92Q.pdf. Authoring organization: NORC at the University of Chicago. Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation Endline Report.

Alicia Menendez, Ursula Hoadley, and Anna Solovyeva. Understanding improvements in reading performance in Liberia: The centrality of text. Unpublished working paper, 2023.

MSI. Quality instruction towards access and basic education improvement (QITABI) endline report, 2018. Sponsoring organization: United States Agency for International Development. Document type: Endline Report.

MSI. RAMP impact evaluation final report, 2019. URL https://pdf.usaid.gov/pdf_docs/PA00WKQ9.pdf. Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation Final Report.

Amy Mulcahy-Dunn and Thomaz Alvares de Azevedo. Ten years of early grade reading programming: Retrospective (2011-2021). USAID Report, February 2023. URL https://www.edu-links.org/sites/default/files/media/file/Ten_Years_of_Early_Grade_Reading_Programming_A_Retrospective.pdf.

National Reading Panel (US) and National Institute of Child Health and Human Development (US). Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups. National Institute of Child Health and Human Development, National Institutes of Health, 2000.

NORC. Performance & impact evaluation (p&ie). final performance evaluation: School health and reading program (shrp), 2016. URL https://pdf.usaid.gov/pdf_docs/PBAAH455.pdf. Sponsoring organization: United States Agency for International Development. Document type: Performance and Impact Evaluation.

NORC. Uganda performance and impact evaluation for literacy achievement and retention activity (lara). midterm impact and final performance evaluation report, 2020. URL https://www.norc.org/PDFs/LARA/LARA%20P%20and%20IE_Midterm%20IE%20and%20Final%20PE%20Report.pdf. Sponsoring organization: United States Agency for International Development. Document type: Performance and Impact Evaluation.

Alejandro Ome, Yvonne Cao, Michel Rousseau, Elise Le, and Russell Owen. Usaid/georgia external impact evaluation of the georgian primary education (g-pried) project. endline impact evaluation report (final), 2016. URL https://pdf.usaid.gov/pdf_docs/pa00m256.pdf. Authoring organization: NORC at the University of Chicago. Sponsoring organization: United States Agency for International Development. Document type: Endline Impact Evaluation Report.

Benjamin Piper and Abel Mugenda. The primary math and reading (primr) initiative. endline impact evaluation, 2014. URL https://pdf.usaid.gov/pdf_docs/PA00K27S.pdf. Authoring organization: RTI International. Sponsoring organization: United States Agency for International Development (USAID)/Kenya. Document type: Endline Impact Evaluation.

Benjamin Piper, Stephanie Simmons Zuilkowski, and Abel Mugenda. Improving reading outcomes in Kenya: First-year effects of the PRIMR Initiative. International Journal of Educational Development, 37:11–21, 2014.

Benjamin Piper, Stephanie Simmons Zuilkowski, Dunston Kwayumba, and Carmen Strigel. 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, 2016.

Benjamin Piper, Joseph Destefano, Esther M Kinyanjui, and Salome Ong'ele. Scaling up successfully: Lessons from Kenya's Tusome national literacy program. Journal of Educational Change, 19:293–321, 2018.

Magda Raupp, Bruce Newman, Luis Revés, Carlos Lauchande, and Edward Jay Allan. Impact evaluation of the usaid/ aprender a ler project in mozambique year 3 ie/rct. final report, 2016. URL https://pdf.usaid.gov/pdf_docs/pa00m5d4.pdf. Authoring organization: International Business & Technical Consultants, Inc. (IBTCI). Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation Final Report.

Chitanda Rhodwell. Read to succeed. endline survey report, 2017. URL https://pdf.usaid.gov/pdf_docs/PA00MW9V.pdf. Authoring organization: Creative Associates International, Inc. Sponsoring organization: United States Agency for International Development (USAID), the USAID/Zambia Read to Succeed Project. Document type: Endline Survey Report.

Jordan Robinson, Mike Duthie, and Andrea Hur. Basa pilipinas impact evaluation final report, 2017. URL https://pdf.usaid.gov/pdf_docs/pa00d5qv.pdf. Authoring organization: Social Impact, Inc.. Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation Final Report.

RTI. Girls' improved learning outcomes (gilo): Final report, 2014. URL https://learningportal.iiep.unesco.org/en/library/girls-improved-learning-outcomes-gilo-final-report. Sponsoring organization: United States Agency for International Development/Egypt. Document type: Final Report.

RTI. Haiti tout timoun ap li - total: Final report, revised, 2015a. URL https://shared.rti.org/content/haiti-tout-timoun-ap-li-total-final-report-revised#. Sponsoring organization: United States Agency for International Development/Haiti. Document type: Final Report.

RTI. Nigeria reading and access research activity(rara). results of an approach to improve early grade reading in hausa in bauchi and sokoto states, 2015b. URL https://shared.rti.org/content/nigeria-reading-and-access-research-activity-rara-results-approach-improve-early-grade#. Sponsoring organization: United States Agency for International Development. Document type: Evaluation.

RTI. Usaid prioritizing reform, innovation, and opportunities for reaching indonesia's teachers, administrators, and students (usaid prioritas). endline monitoring report, volume 3: An assessment of early grade reading - how well children are reading in usaid prioritas districts (cohorts 1, 2, and 3), 2017. URL https://pdf.usaid.gov/pdf_docs/PA00N1N6.pdf. Sponsoring organization: United States Agency for International Development. Document type: Endline Report.

RTI. Assistance to basic education: All children reading (abe acr). merit: The malawi early grade reading improvement activity national assessment of reading instruction, standards 1-4, 2019. URL https://pdf.usaid.gov/pdf_docs/PA00XKFT.pdf. Sponsoring organization: United States Agency for International Development. Document type: Final Project Report.

RTI and CNEM. Evaluation initiale des competences fondamentales en lecture-ecriture basee sur l'utilisation de l'outil egra adapte en français et en arabe au mali, 2009. URL https://pdf.usaid.gov/pdf_docs/PA00J2TM.pdf. Sponsoring organization: United States Agency for International Development. Document type: Evaluation.

Save the Children. Read bangladesh: Final report, 2018. URL https://pdf.usaid.gov/pdf_docs/PA00TJ2D.pdf. Sponsoring organization: United States Agency for International Development. Document type: Final Report.

School-to-School. Tz21 endline evaluation report: Evaluation conducted in zanzibar, mtwara, and lindi, 2015. URL https://sts-international.org/wp-content/uploads/2019/09/wherewework_tanzania21_content1.pdf. Sponsoring organization: United States Agency for International Development. Document type: Endline Evaluation Report.

Robert E Slavin and Alan Cheung. A synthesis of research on language of reading instruction for english language learners. Review of Educational Research, 75(2):247–284, 2005.

Social Impact. Impact evaluation of the early grade reading activity (egra). final report, 2018. URL https://pdf.usaid.gov/pdf_docs/PA00T3Q6.pdf. Sponsoring organization: United States Agency for International Development. Document type: Impact Evaluation.

Thomas Tilson, Augustine Kamlongera, Mateusz Pucilowski, and Dr. Dorothy Nampota. Evaluation of the malawi teacher professional development support (mtpds). program final evaluation report, 2013. URL https://pdf.usaid.gov/pdf_docs/pdacx458.pdf. Authoring organization: Social Impact, Inc. (SI). Sponsoring organization: United States Agency for International Development. Document type: Final Evaluation Report.

Elizabeth Tipton. Small sample adjustments for robust variance estimation with meta-regression. Psychological methods, 20(3):375, 2015.

Adam Turney, David Noyes, Flávio Magaia, Trymore Mafucha Dhliwayo, Yuri Machkasov, Euclides Chigogolo Dinis Zacarias, and Haiyan Hua. Effectiveness evaluation early grade reading assessment & supplementary tools. midline report - mozambique, 2020. URL https://pdf.usaid.gov/pdf_docs/PA00X21D.pdf. Authoring organization: Creative Associates International, World Education, Inc.. Sponsoring organization: United States Agency for International Development. Document type: Effectiveness Evaluation Midline Report.

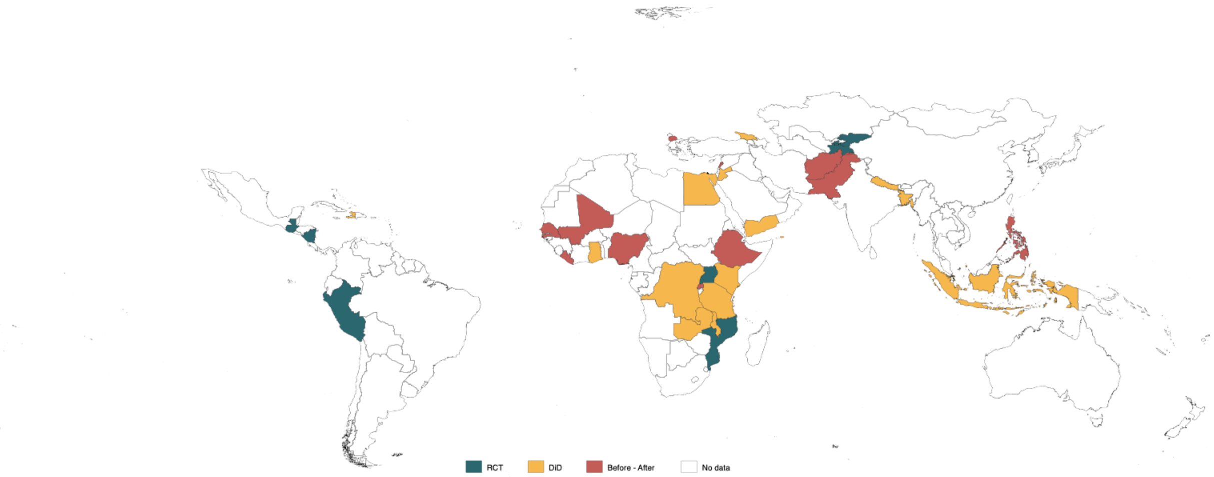
Tyler J VanderWeele. Mediation analysis: a practitioner's guide. Annual review of public health, 37:17–32, 2016.

Wolfgang Viechtbauer, José Antonio López-López, Julio Sánchez-Meca, and Fulgencio Marín-Martínez. A comparison of procedures to test for moderators in mixed-effects meta-regression models. Psychological Methods, 20(3), 2015.

Elena Vinogradova and Gabriel Montero. Whole school reading program evaluation findings, 2013. URL https://pdf.usaid.gov/pdf_docs/PA00M37Q.pdf. Authoring organization:

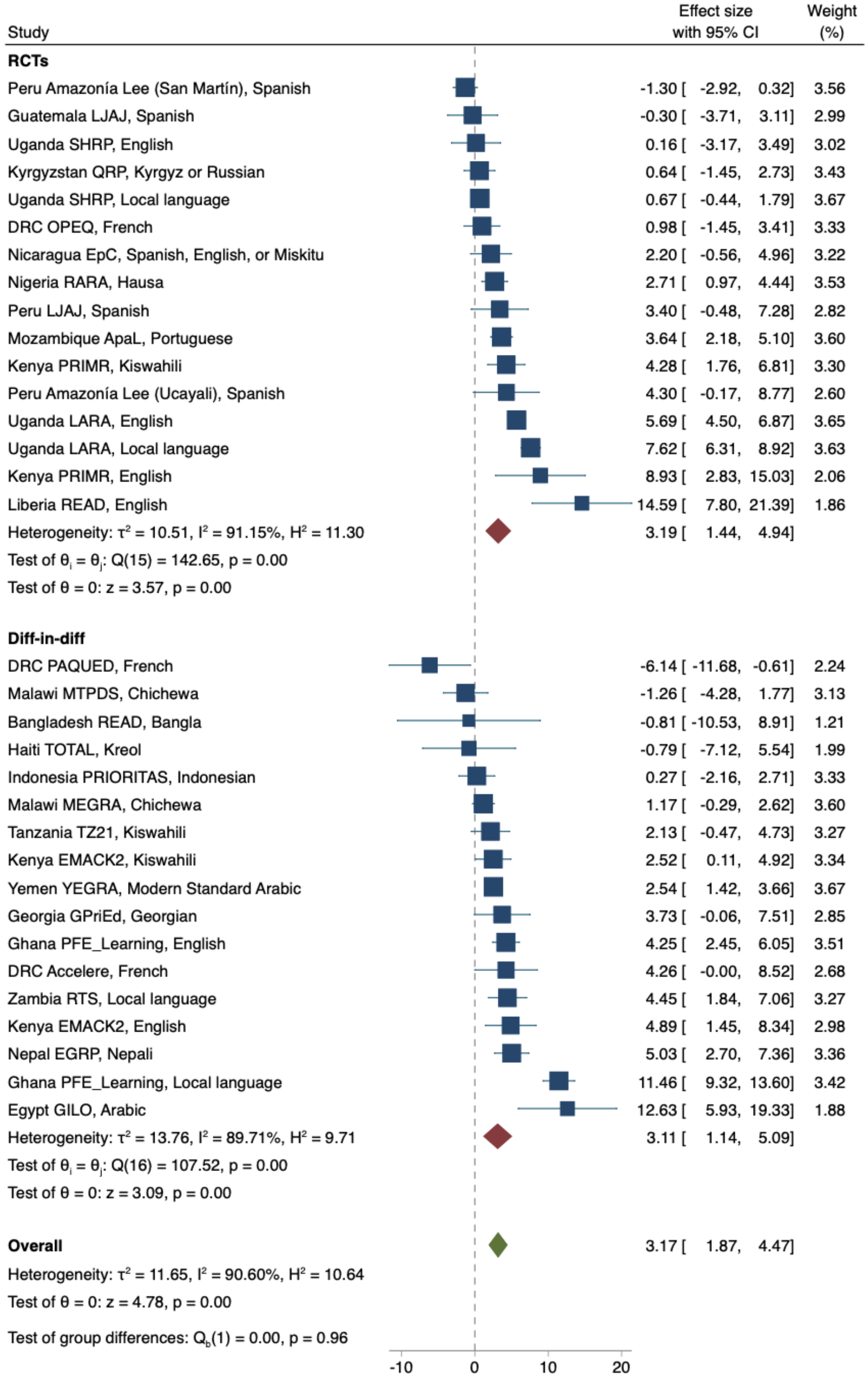
Education Development Center Inc.. Sponsoring organization: United States Agency for International Development/Philippines. Document type: Impact Evaluation Final Report.

Figure 1: Map of early-grade reading evaluations by methodology



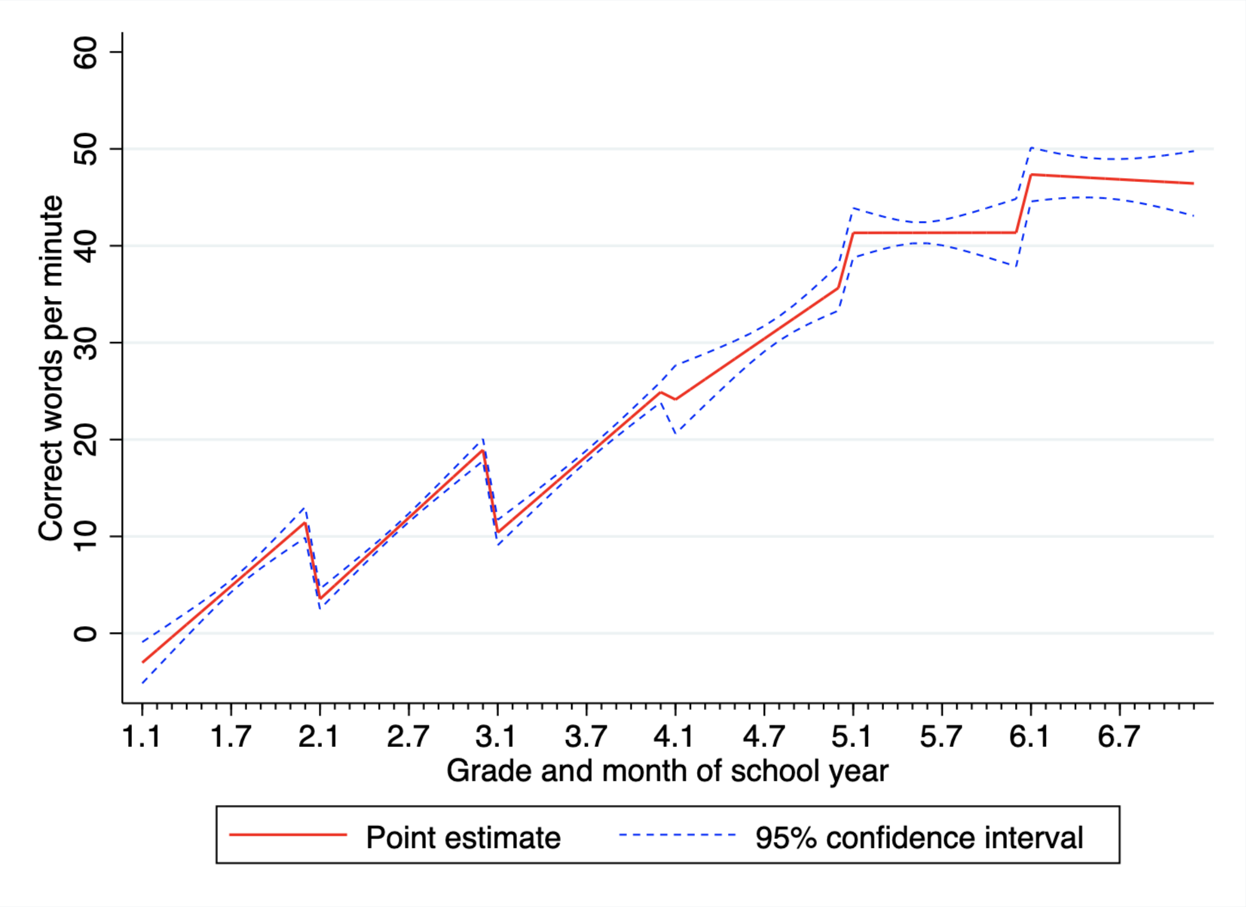
Note: Coverage includes only evaluations for which microdata is available for the analysis here. When more than one evaluation is available in a given country, the country is coded according to the most rigorous evaluation available (i.e., RCTs take precedence over diff-in-diff evaluations which take precedence over before-and-after evaluations).

Figure 2: Impacts on oral ready fluency



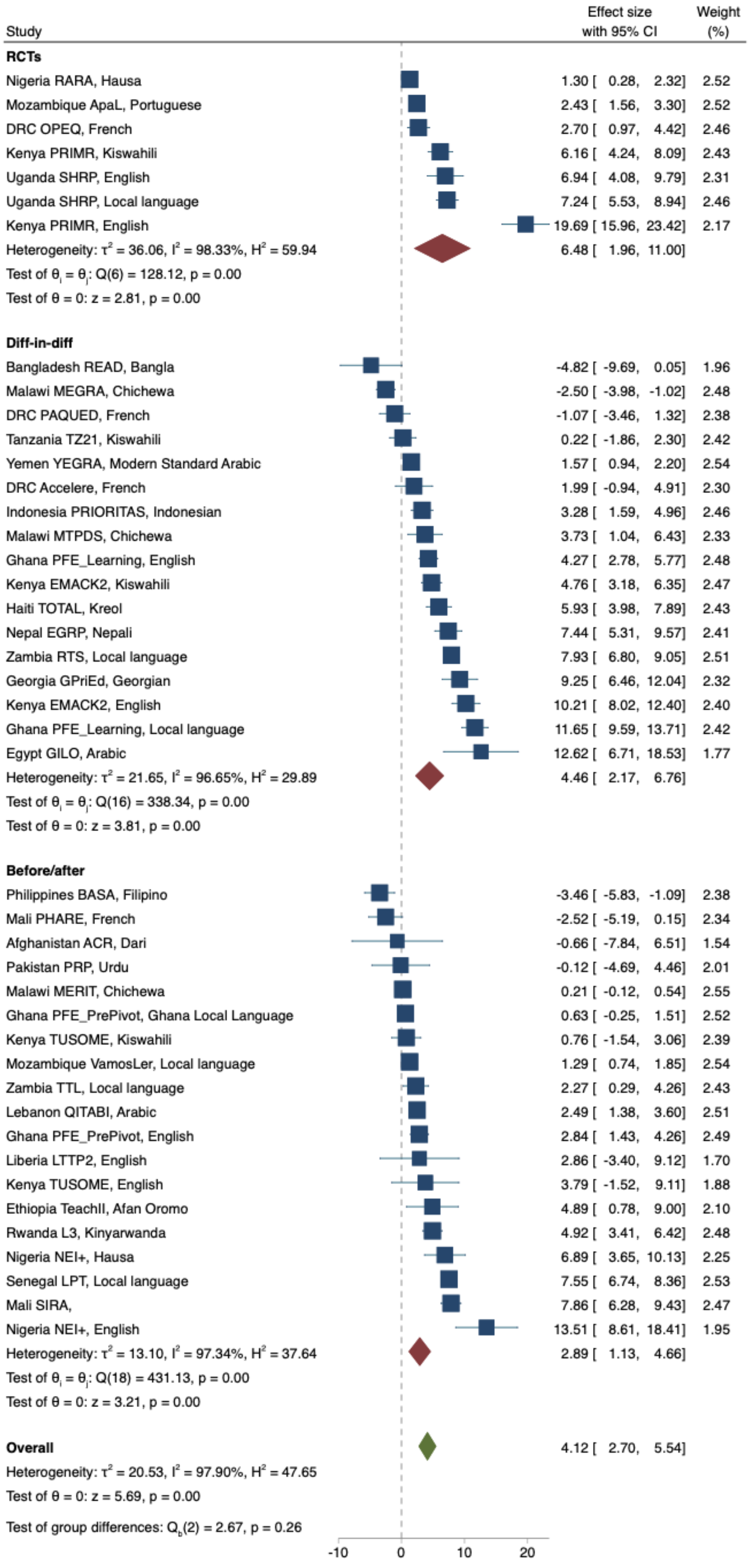
Note: Each coefficient represents an estimated treatment effect, based on the author's reanalysis of microdata. The outcome for all studies, oral reading fluency, is expressed in correct words per minute (CWPM). Overall effects for each group represent empirical Bayes estimates of a random effects model.

Figure 3: Average oral reading fluency by grade and month, pooling all studies



Note: Estimates are based on a pooled regression of correct words per minute (CWPM) on grade dummies, month of school year dummies, and their interaction, as well as controls for study fixed effects and treatment effects, as shown in equation (5).

Figure 4: Changes in oral reading fluency in the treatment group (only)



Note: The outcome for all studies, oral reading fluency, is expressed in correct words per minute (CWPM). Each coefficient represents a first-difference in oral reading fluency, i.e., the change over time in the average reading level for pupils in a given grade (rather than for a cohort progressing from grade to grade over time). Outcomes are adjusted for the timing of data collection when this differs across survey rounds, per Figure 3. Averages for each group represent empirical Bayes estimates of a random effects model.

Figure 5: Average unit cost versus scale

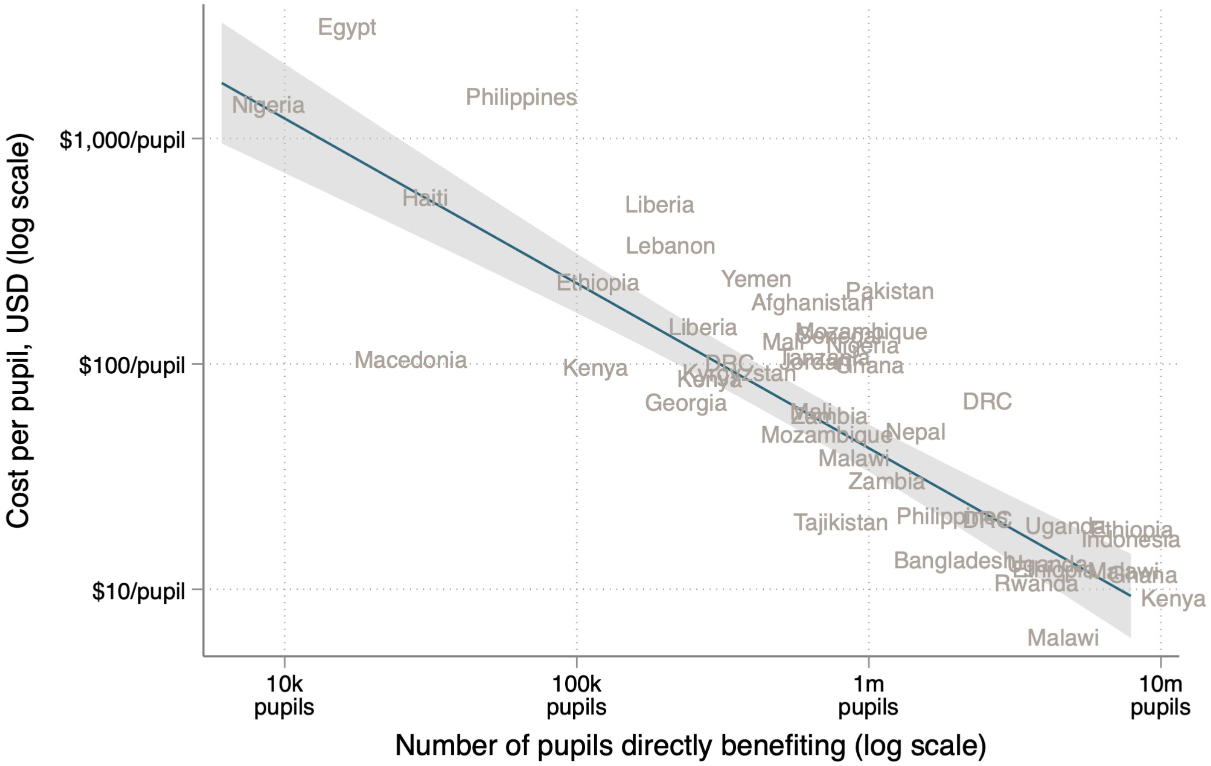
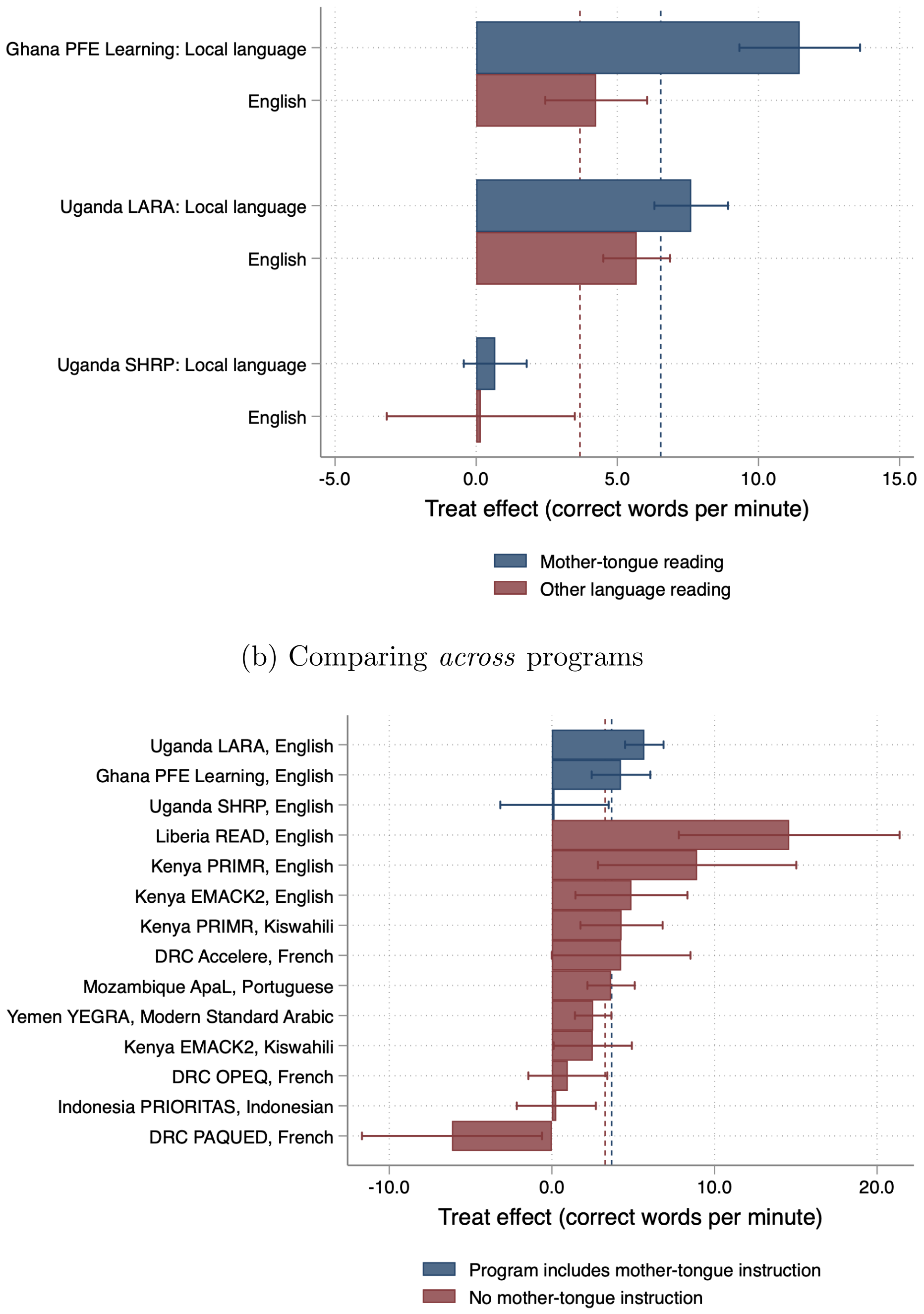


Figure 6: The impact of mother-tongue instruction



(a) Comparing *within* programs

(b) Comparing *across* programs

Note: Dotted lines show average effects for each group based on empirical Bayes estimates of a random effects model.

Table 1: Project meta-data

Country/Project	Grade	Dates	Budget, USD (millions)	Pupils treated (thousands)	Original evaluation reference
RCTs					
DRC - OPEQ	2 to 5	2011/12 - 2013/14	$25.9	253	Aber et al. (2015)
Guatemala - LJAJ	1 to 3	2013/14 - 2015/16			Lugo-Gil et al. (2019a)
Kenya - PRIMR	1 and 2	2012 - 2014	$8.1	83	Piper and Mugenda (2014)
Kyrgyz Republic - QRP	1 to 4	201415 - 2016/17	$19.7	212	AIR and Save the Children (2017a)
Liberia - READ	K1 to 2	2018 - 2021	$28.2	190	Menendez et al. (2021)
Mozambique - ApaL	2 and 3	2013 - 2016	$19.5	395	Raupp et al. (2016)
Nicaragua - EpC	1 to 3	2014/15 - 2016/17			Bagby et al. (2019)
Nigeria - RARA	1 and 2	2014/15 - 2014/15	$8.7	6	RTI (2015b)
Peru - Amazonía Lee	2	2015 - 2016			Campuzano et al. (2018)
Peru - LJAJ	1 to 3	2013/14 - 2015/16			Lugo-Gil et al. (2019b)
Tajikistan - QRP	1 to 4	201415 - 2016/17	$10.3	512	AIR and Save the Children (2017b)
Uganda - LARA	1 to 4	2016 - 2020	$36	2740	NORC (2020)
Uganda - SHRP	1 to 4	2013 - 2019	$61.5	3169	NORC (2016)
Non-randomized diff-in-diff					
Bangladesh - READ	1 to 3	2014 - 2018	$15.4	1124	Save the Children (2018)
DRC - Accelere	1 to 4	2016/17 - 2019/20	$134	1938	IBTCI (2020)
DRC - PAQUED	2, 4 and 6	2010 - 2014	$40	1934	Iriondo-Perez et al. (2014)
Egypt - GILO	1 to 3	2009 - 2011	$38	12	RTI (2014)
Georgia - GPRIED	1 to 6	2013/14 - 2016/17	$10.8	158	Ome et al. (2016)
Ghana - PFE Learning	K2 to 2	2017/18 - 2018/19	$71.0	708	Evaluating Systems (2019)
Haiti - TOTAL	1 and 2	2012/13 - 2013/14	$12.9	23	RTI (2015a)
Indonesia - PRIORITAS	1 to 6	2013/14 - 2016/17	$83.7	4920	RTI (2017)
Jordan - RAMP	K2 to 3	2016/17 - 2018/19	$47.8	458	MSI (2019)
Kenya - EMACK2	1 to 3	2013 - 2014	$17.8	204	Adelman et al. (2014)
Malawi - MEGRA	1 to 3	2013/14 - 2015/16	$24	620	Social Impact (2018)
Malawi - MTPDS	1 to 4	2010/11 - 2012/13	$20	3232	Tilson et al. (2013)
Nepal - EGRP	1 to 3	2016/17 - 2019/20	$53.8	1055	Menendez and Haugan (2018)
Philippines - EQUALLS 2	1 to 6	2012/13 - 2012/13	$60	39	Vinogradova and Montero (2013)
Tanzania - TZ21	1 to 4	2013 - 2015	$48.9	451	School-to-School (2015)
Yemen - CLP	1 to 3	2012/13 - 2014/15	$70	288	Baraheem (2013)
Zambia - RTS	1 to 3	2013 - 2016	$24.1	787	Rhodwell (2017)
Before-and-after					
Afghanistan - ACR	1 to 3	2017/18 - 2020/21	$69.5	366	Creative Associates (2019)
Ethiopia - IQPEP	1 to 8	2010/11 - 2013/14	$35.6	2859	IQPEP (2014)
Ethiopia - READ	1 to 8	2014/15 - 2018/19	$99	5287	AIR (2019)
Ethiopia - Teach II	1 to 4	2012/13 - 2014/15	$18.3	79	Gebrekidan (2014)
Ghana - PFE PrePivot	K1 to 3	2013/14 - 2016/17	$71.8	6106	Ghana Education Service et al. (2016)
Kenya - Tusome	1 and 2	2015 - 2019	$73.1	7891	Keaveney et al. (2020)
Lebanon - QITABI	1 to 4	2015/16 - 2018/19	$46.3	136	MSI (2018)
Liberia - LTTP2	1 to 3	2010 - 2015	$70	135	King et al. (2015)
Macedonia - RAL	1 to 5	2014/15 - 2016/17	$1.7	16	
Malawi - MERIT	1 to 4	2016/17 - 2019/20	$63.5	5180	RTI (2019)
Mali - PHARE	1 to 6	2009/10 - 2011/12	$31	497	RTI and CNEM (2009)
Mali - SIRA	1 and 2	2017/18 - 2020/21	$51	398	
Mozambique - Vamos Ler	1 to 4	2018 - 2021	$73.5	518	Turney et al. (2020)
Nigeria - NEI+	1 to 3	2016 - 2020	$81	659	Campos et al. (2017)
Pakistan - PRP	1 and 2	2015/16 - 2019/20	$164.7	769	Koch et al. (2018)
Philippines - BASA	1 to 3	2013/14 - 2015/16	$24.7	1149	Robinson et al. (2017)
Rwanda - L3	1 to 4	2012 - 2016	$26.8	2485	Education Development Center (2017)
Senegal - LPT	1 and 2	2017/18 - 2020/21	$71	524	EdIntersect and Chemonics (2021)
Zambia - TTL	1 to 4	2013 - 2016	$30	501	Falconer-Stout et al. (2017)

Note: Whenever available, the benchmark means and standard deviations are from the control group at baseline. Where no baseline is available, the table reports the value for the control group. Conversely, where no control group is available, it reports the baseline value for the treatment group. Baseline data for Uganda SHRP and Kyrgyzstan QRP is not available, so balance tests are omitted. The number of pupils treated is measured in an intent-to-treat sense, i.e., pupils in schools and grades assigned to treatment.

Table 2: Intervention variants

	Training teachers with evidence-based curriculum	Providing instructional guidelines	Following up with coaching and monitoring	Providing instructional materials	Providing tools and training for student assessment	Mother-tongue instruction	
Afghanistan	ACR	x	x	x	x	x	
Bangladesh	READ	x		x	x	x	
DRC	Accelere	x	x	x	x	x	
DRC	OPEQ	x	x	x	x		
DRC	PAQUED	x	x	x	x		
Egypt	GILO	x	x	x	x	x	
Ethiopia	IQPEP	x	x				x
Ethiopia	READ	x		x	x	x	x
Ethiopia	TeachII	x			x		x
Georgia	GPriEd	x		x	x	x	
Ghana	PFE Learning	x		x	x		x
Ghana	PFE PrePivot	x	x	x	x	x	x
Guatemala	LJAJ	x	x	x	x	x	x
Haiti	TOTAL	x		x	x	x	
Indonesia	PRIORITAS	x		x	x	x	
Jordan	RAMP	x		x	x	x	
Kenya	EMACK2	x		x			
Kenya	PRIMR	x		x	x		
Kenya	Tusome	x	x	x	x	x	
Kyrgyzstan	QRP	x		x	x		
Lebanon	QITABI	x		x	x		
Liberia	LTTP2	x	x	x	x	x	
Liberia	READ	x	x	x	x	x	
Macedonia	RAL	x	x	x		x	
Malawi	MEGRA	x	x	x	x		
Malawi	MERIT	x		x	x		
Malawi	MTPDS	x	x	x	x		
Mali	PHARE	x		x	x	x	
Mali	SIRA	x	x	x	x	x	
Mozambique	ApaL	x	x		x	x	
Mozambique	VamosLer	x		x	x		x
Nepal	EGRP	x	x	x	x	x	
Nicaragua	EpC	x		x	x	x	
Nigeria	NEI+	x	x	x	x		x
Nigeria	RARA	x	x	x	x		x
Pakistan	PRP	x	x				x
Peru	Amazonía Lee	x		x	x	x	
Peru	LJAJ	x	x	x	x	x	x
Philippines	BASA	x	x	x	x		
Philippines	EQuALLS2	x		x	x	x	
Rwanda	L3	x		x	x	x	
Senegal	LPT	x	x	x	x	x	x
Tajikistan	QRP	x		x	x	x	
Tanzania	TZ21	x		x	x		
Uganda	LARA	x	x	x	x	x	x
Uganda	SHRP	x	x	x	x		x
Yemen	YEGRA	x	x	x	x	x	
Zambia	RTS	x		x	x	x	x
Zambia	TTL	x		x	x	x	x

Source: coded on the basis of project reports listed in Table 1 and Graham and Kelly (2019).

Table 3: Summary statistics & balance

		Baseline and/or control group: correct words per minute			Balance test	
USAID Project	Language	% Zero	Mean	Std. Dev.	T minus C	p-value
RCTs						
Uganda SHRP	Local language	95.8	0.5	3.2	0.2	0.48
Nigeria RARA	Hausa	94.0	1.1	5.1	-0.7	0.11
Uganda SHRP	English	91.2	0.7	3.5	0.2	0.55
Mozambique ApaL	Portuguese	80.4	1.6	10.4	-0.7	0.18
DRC OPEQ	French	68.9	6.4	12.7	-0.1	0.98
Kenya PRIMR	Kiswahili	49.8	11.8	16.0	-0.6	0.77
Kenya PRIMR	English	43.6	16.8	23.2	-1.0	0.77
Uganda LARA	Local language	42.6	10.2	11.9		
Liberia READ	English	33.8	15.0	18.6	-1.1	0.53
Uganda LARA	English	22.5	21.1	19.5		
Kyrgyzstan QRP	Kyrgyz or Russian	1.2	47.9	28.5		
Peru LJAJ	Spanish					
Peru Amazonía Lee (Ucayali)	Spanish					
Nicaragua EpC	Spanish, English, or Miskitu					
Guatemala LJAJ	Spanish					
Peru Amazonía Lee (San Martín)	Spanish					
Non-randomized diff-in-diff						
Malawi MTPDS	Chichewa	90.8	1.3	5.8	-0.2	0.80
Yemen YEGRA	Modern Standard Arabic	85.5	0.6	2.3	-0.0	0.90
Ghana PFE Learning	Local language	85.0	1.4	5.7	0.6	0.12
Zambia RTS	Local language	77.5	3.8	9.3	-2.1	0.02
Nepal EGRP	Nepali	56.6	9.8	16.8	0.0	0.99
Egypt GILO	Arabic	51.5	9.0	14.7	2.7	0.15
DRC Accelere	French	51.4	11.9	17.7	4.2	0.01
DRC PAQUED	French	50.5	13.8	18.6	2.1	0.18
Ghana PFE Learning	English	50.4	5.4	11.4	0.5	0.63
Haiti TOTAL	Kreol	48.8	8.5	16.0	-1.6	0.37
Malawi MEGRA	Chichewa	48.4	14.6	18.3	0.6	0.48
Kenya EMACK2	English	47.9	19.1	26.3	0.3	0.88
Kenya EMACK2	Kiswahili	45.6	17.3	20.0	1.7	0.24
Tanzania TZ21	Kiswahili	44.7	10.9	13.1	-2.7	0.25
Bangladesh READ	Bangla	29.5	22.0	28.2	3.6	0.31
Indonesia PRIORITAS	Indonesian	3.4	58.8	28.5	5.0	0.03
Georgia GPriEd	Georgian	0.3	54.5	31.2	3.6	0.20
Before-and-after						
Mozambique VamosLer	Three languages*	95.0	0.4	2.3		
Zambia TTL	Local language	92.0	1.1	4.5		
Nigeria NEI+	Hausa	80.2	3.5	10.6		
Nigeria NEI+	English	79.7	2.0	5.9		
Ghana PFE PrePivot	Local language	79.7	3.3	9.7		
Mali PHARE	French	73.1	9.1	17.1		
Senegal LPT	Three languages*	70.1	1.7	4.2		
Mali SIRA		68.2	3.5	9.3		
Malawi MERIT	Chichewa	65.9	6.6	12.1		
Ghana PFE PrePivot	English	52.4	8.7	17.4		
Liberia LTTP2	English	46.6	12.9	19.0		
Kenya TUSOME	Kiswahili	44.4	13.3	16.2		
Ethiopia TeachII	Afan Oromo	39.9	16.4	20.5		
Rwanda L3	Kinyarwanda	36.6	15.9	16.7		
Afghanistan ACR	Dari	34.0	26.3	40.0		
Kenya TUSOME	English	32.2	26.7	29.7		
Pakistan PRP	Urdu	21.4	38.3	32.6		
Philippines BASA	Filipino	6.2	35.5	20.6		
Lebanon QITABI	Arabic	5.7	23.0	17.3		

Note: Whenever available, the benchmark means and standard deviations are from the control group at baseline. Where no baseline is available, the table reports the value for the control group (and where no control group is available, it reports the baseline value for the treatment group). Baseline data for Uganda SHRP and Kyrgyzstan QRP is not available, so balance tests are omitted.

* In Mozambique, testing was conducted in either Echuwabo, Elomwe, or Emakhuwa, while in Senegal testing was done in either Wolof, Pulaar, or Serere.

Table 4: Treatment effects on oral reading fluency

Country / project	Language	Coefficient	Standard Error	N (Pupils)	N (Clusters)	Treated clusters	School FE
RCTs							
DRC OPEQ	French	1.0	[1.2]	6248	76	48	Yes
Guatemala LJAJ	Spanish	-0.3	[1.7]	880			
Kenya PRIMR	English	8.9	[3.1]	8592	32	24	Yes
Kenya PRIMR	Kiswahili	4.3	[1.2]	8589	32	24	Yes
Kyrgyzstan QRP	Kyrgyz or Russian	0.6	[1.0]	5066	131	65	
Liberia READ	English	14.6	[3.4]	2692	88	43	
Mozambique ApaL	Portuguese	3.6	[0.7]	3433	2264	1129	
Nicaragua EpC	Spanish, English, or Miskitu	2.2	[1.4]	2349			
Nigeria RARA	Hausa	2.7	[0.8]	2573	120	60	Yes
Peru Amazonía Lee (San Martín)	Spanish	-1.3	[0.8]	1646			
Peru Amazonía Lee (Ucayali)	Spanish	4.3	[2.2]	740			
Peru LJAJ	Spanish	3.4	[1.9]	684			
Uganda LARA	English	5.7	[0.6]	4910	20	14	
Uganda LARA	Local language	7.6	[0.6]	4912	20	14	
Uganda SHRP	English	0.2	[1.6]	28682	96	55	Yes
Uganda SHRP	Local language	0.7	[0.5]	32952	96	55	Yes
Non-randomized diff-in-diff							
Bangladesh READ	Bangla	-0.8	[4.9]	2127	37	20	Yes
DRC Accelere	French	4.3	[2.1]	3885	236	118	Yes
DRC PAQUED	French	-6.1	[2.8]	7048	145	111	
Egypt GILO	Arabic	12.6	[3.4]	2232	56	28	Yes
Georgia GPriEd	Georgian	3.7	[1.9]	4163	203	102	Yes
Ghana PFE Learning	English	4.3	[0.9]	17831	139	81	
Ghana PFE Learning	Local language	11.5	[1.0]	17830	139	81	
Haiti TOTAL	Kreol	-0.8	[3.2]	2744	57	28	Yes
Indonesia PRIORITAS	Indonesian	0.3	[1.2]	15087	341	172	Yes
Kenya EMACK2	English	4.9	[1.7]	5383	95	50	Yes
Kenya EMACK2	Kiswahili	2.5	[1.2]	5381	95	50	Yes
Malawi MEGRA	Chichewa	1.2	[0.7]	18462	10	10	
Malawi MTPDS	Chichewa	-1.3	[1.5]	1913	51	21	
Nepal EGRP	Nepali	5.0	[1.1]	10619	205	86	Yes
Tanzania TZ21	Kiswahili	2.1	[1.3]	2334	19	16	
Yemen YEGRA	Modern Standard Arabic	2.5	[0.5]	3556	88	43	
Zambia RTS	Local language	4.4	[1.3]	7923	95	78	

Table 5: Meta-regressions

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.936***	8.740*	0.305	2.981***	-0.066	5.271
	(0.507)	(4.857)	(1.825)	(0.909)	(1.443)	(13.745)
Log no. of beneficiaries		-0.986				-1.055
		(0.825)				(1.821)
Log cost per pupil (USD)			1.515			0.289
			(1.030)			(2.553)
RCT==1				0.116		0.558
				(1.252)		(1.213)
Diff-in-diff==1				-0.236		
				(1.279)		
Starting year (2010=1)					0.630**	0.678**
					(0.298)	(0.339)
N (Estimates)	52	47	47	52	52	47
N (Studies)	43	43	43	43	43	43

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	2.901***	10.326	-0.856	2.756***	-1.232	-4.415
	(0.600)	(6.675)	(2.766)	(0.889)	(1.898)	(16.725)
Log no. of beneficiaries		-1.277				-0.063
		(1.166)				(2.259)
Log cost per pupil (USD)			2.163			1.936
			(1.501)			(3.266)
RCT==1				0.316		0.109
				(1.227)		(1.048)
Starting year (2010=1)					0.926**	0.979**
					(0.384)	(0.430)
N (Estimates)	33	28	28	33	33	28
N (Studies)	27	27	27	27	27	27

Note: The table reports results from meta-analytic regressions. The dependent variable is the coefficient on the treatment effect, with one observation per study and outcome language. The specification allows for correlated random effects at the study level.

Table 6: Intermediate outcomes

Liberia READ

		Does your teacher. . .			
	Do you have books at school that you can take home to read?	make you practice silent reading in class?	make you practice reading out loud in class?	assign reading for you to do at home?	ever make you re-tell a story during class?
	(1)	(2)	(3)	(4)	(5)
Treatment	0.195***	0.113	0.126*	0.178**	0.161**
	(0.045)	(0.075)	(0.070)	(0.080)	(0.076)
Constant	0.550***	1.663***	1.983***	1.442***	1.245***
	(0.036)	(0.052)	(0.045)	(0.055)	(0.051)
N (Pupils)	1,200	1,200	1,200	1,200	1,200

Uganda SHRP

	Teacher guides students to read words from printed material?	Teacher uses textbook?	Teacher taught lesson in official school language?	How many learners have textbook or printed material?	Does the teacher have any records of learner assessment?
	(1)	(2)	(3)	(4)	(5)
Treatment	1.000***	0.940***	0.304***	0.918***	0.611***
	(0.000)	(0.060)	(0.091)	(0.065)	(0.133)
Constant	-0.000**	0.060	0.678***	0.001	0.031
	(0.000)	(0.059)	(0.089)	(0.003)	(0.028)
N (Pupils)	975	975	975	975	975

Kenya PRIMR

	Number of times observed:				
	Reading comprehension	Reading out loud	Silent reading	Letters and sounds	Textbook
	(1)	(2)	(3)	(4)	(5)
Treatment	2.019***	1.607***	0.312***	2.266***	3.103***
	(0.343)	(0.196)	(0.066)	(0.273)	(0.640)
Constant	0.628**	0.498***	0.011	0.297	3.057***
	(0.289)	(0.122)	(0.012)	(0.213)	(0.559)
N (Pupils)	4,222	4,222	4,222	4,222	4,222

Kyrgyzstan QRP

	Do you have a reading textbook?	Do you get reading homework?	Teacher check your reading skills in past 7 days?	Reading activity activity outside regular class?
	(1)	(2)	(3)	(4)
Treatment	-0.002	-0.000	0.018	0.009
	(0.004)	(0.002)	(0.011)	(0.029)
Constant	0.989***	0.995***	0.280***	0.547***
	(0.002)	(0.001)	(0.007)	(0.020)
N (Pupils)	5,084	5,003	5,003	5,023

Table 7: Associations between intermediate outcomes and learning gains

Liberia READ

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Do you have books at school that you can take home to read?	3.816**						0.319
	(1.784)						(1.629)
Does your teacher make you practice silent reading in class?		1.923**					0.729
		(0.904)					(0.899)
Does your teacher make you practice reading out loud in class?			2.748***				1.872*
			(0.957)				(1.021)
Does your teacher assign reading for you to do at home?				1.292			-0.064
				(1.148)			(1.043)
Does your teacher ever make you re-tell a story during class?					1.597		0.474
					(0.986)		(0.903)
Treatment						13.511***	13.066***
						(3.019)	(2.980)
Constant	22.119***	21.264***	18.951***	22.591***	22.454***	18.279***	12.682***
	(1.828)	(2.044)	(2.205)	(2.130)	(1.983)	(1.632)	(2.966)
N (Pupils)	1,200	1,200	1,200	1,200	1,200	1,200	1,200

Uganda SHRP

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher guides students to read words from printed material?	2.944						31.415***
	(2.178)						(1.365)
Teacher uses textbook?		2.838					-1.016
		(2.062)					(1.097)
Teacher taught lesson in official school language?			6.260				2.652
			(3.889)				(1.601)
How many learners have textbook or printed material?				1.220			-31.346***
				(1.310)			(1.785)
Does the teacher have any records of learner assessment?					4.807		0.757
					(2.837)		(0.623)
Treatment						2.944	
						(2.178)	
Constant	0.904**	0.897**	-2.726	1.977*	0.761**	0.904**	-0.836
	(0.402)	(0.426)	(2.098)	(1.127)	(0.278)	(0.402)	(0.914)
N (Pupils)	975	975	975	975	975	975	975

Kenya PRIMR

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No. of times observed: Reading comprehension	2.275***						1.211*
	(0.679)						(0.706)
No. of times observed: Reading out loud		2.189**					0.730
		(0.897)					(1.061)
No. of times observed: Silent reading			-0.142				-4.050*
			(2.068)				(2.366)
No. of times observed: Letters and sounds				0.161			-1.762*
				(0.726)			(1.026)
No. of times observed: Textbook					0.697		-0.232
					(0.433)		(0.436)
Treatment						14.166***	16.524***
						(2.911)	(4.685)
Constant	36.832***	37.988***	41.788***	41.428***	37.975***	30.985***	31.138***
	(1.967)	(2.018)	(1.607)	(2.038)	(2.899)	(2.371)	(3.007)
N (Pupils)	4,222	4,222	4,222	4,222	4,222	4,222	4,222

Kyrgyzstan QRP

	(1)	(2)	(3)	(4)	(5)	(6)
Do you have a reading textbook?	8.826***					8.317**
	(3.359)					(3.495)
Do you get reading homework?		4.114				3.388
		(3.818)				(3.495)
Teacher check your reading skills in past 7 days?			-6.115***			-5.389***
			(1.735)			(1.799)
Reading activity outside regular class?				-6.352***		-6.376***
				(1.292)		(1.268)
Treatment					1.779*	2.009**
					(1.020)	(0.999)
Constant	40.580***	44.999***	51.055***	52.926***	48.335***	41.893***
	(3.377)	(3.872)	(0.710)	(0.849)	(0.699)	(5.002)
N (Pupils)	5,044	4,963	4,963	4,984	5,066	4,801

Table 8: Meta-analysis results for average treatment effects

	Random effects	Fixed effects	Weighted by beneficiaries
Correct words per minute			
RCTs	3.061***	3.481***	3.436
	(0.840)	(1.075)	(2.856)
Diff-in-diff	2.752***	2.818***	0.943
	(0.893)	(0.714)	(1.038)
Total	2.901***	3.139***	1.978
	(0.600)	(0.639)	(1.207)
I^2	88.845	87.210	
Correct letters per minute			
RCTs	4.587***	3.709***	5.081**
	(1.510)	(0.974)	(2.065)
Diff-in-diff	4.318**	4.836***	1.300
	(1.895)	(1.738)	(1.885)
Total	4.525***	4.289***	2.911*
	(1.288)	(1.025)	(1.564)
I^2	95.638	93.059	

Table 9: Incorporating before-and-after studies into overall treatment effect estimate

	First difference, treatment group	First difference, control group	Treatment effect
Correct words per minute			
RCTs + DiD	4.278***	1.453**	2.901***
	(0.973)	(0.618)	(0.600)
Before and after	2.684***		1.093
	(0.946)		(0.909)
Total	3.585***		2.220***
	(0.684)		(0.535)

Note: Numbers in red are imputed, on the basis of the assumption that learning progress in control schools is the same in before-and-after studies as in the (pooled estimate from) the RCTs and diff-in-diff studies.

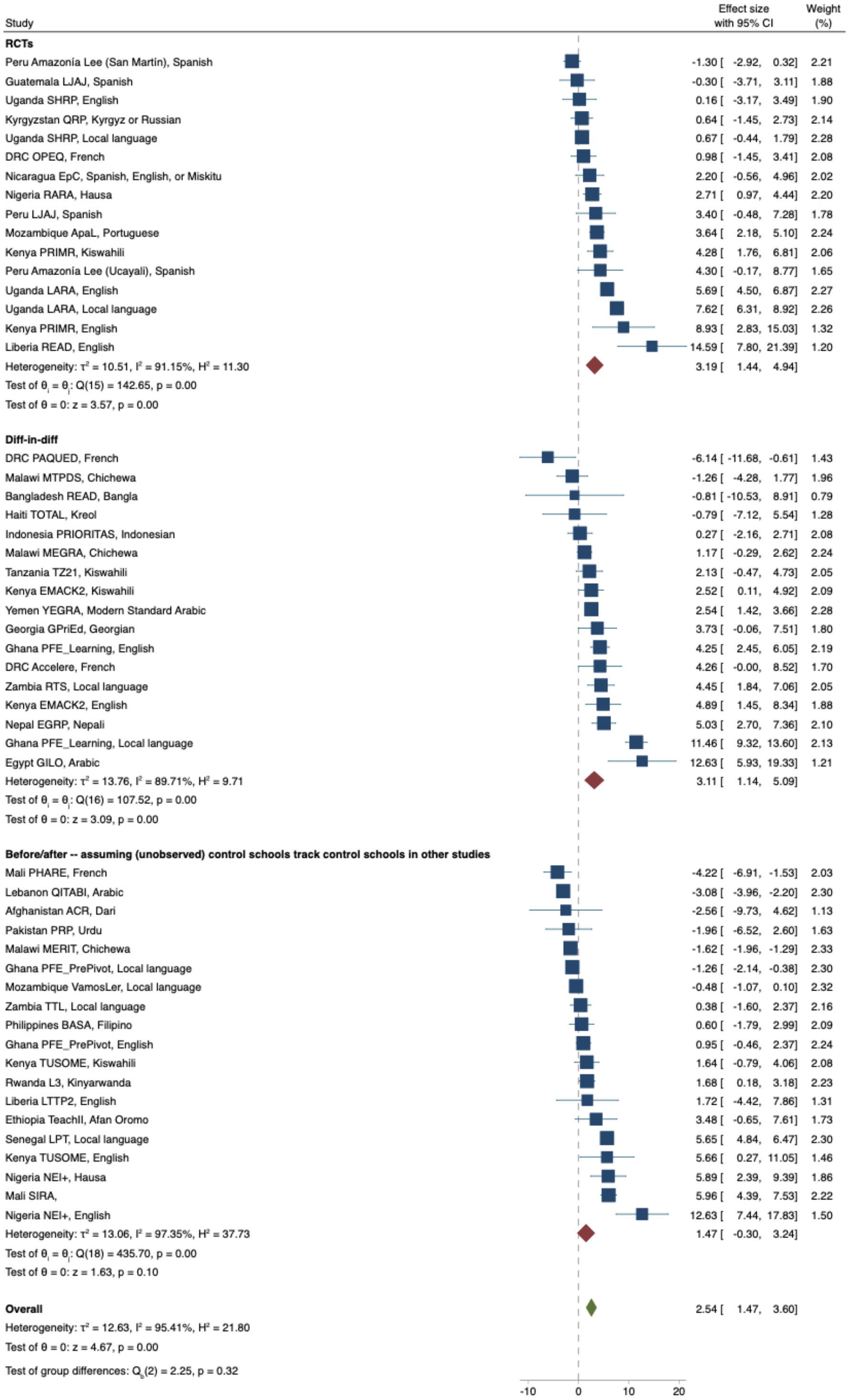
Appendix: Additional table and figures

Table 10: Timing of data collection and school calendar by program

	Baseline				Endline			
	Grade	Test year	Test month	School start month	Grade	Test year	Test month	School start month
RCTs								
DRC OPEQ	3,4	2011	Mar - May	Sep	3,4,5	2013	Apr - Jun	Sep
Kenya PRIMR	1,2	2012	Jan	Jan	1,2	2013	Oct	Jan
Kyrgyzstan QRP					2,4	2017	Apr	Sep
Liberia READ	2	2017	May	Sep	3,4	2021	Mar - Apr	Dec
Mozambique ApaL	2,3	2013	Feb - Mar	Jan	2,3	2015	Sep - Oct	Jan
Nigeria RARA	2	2014	Nov	Sep	2	2015	Jun - Aug	Sep
Uganda LARA	1	2017	Feb	Feb	3	2020	Jul - Oct	Feb
Uganda SHRP	1,3	2013	Jun	Jan	2,3,4	2017	Jun	Feb
Non-randomized diff-in-diff								
Bangladesh READ	2,3	2015	Jun - Jul	Jan	2,3	2016	Nov - Dec	Jan
DRC Accelere	5	2015	Oct - Nov	Sep	5	2019	Oct - Dec	Sep
DRC PAQUED	2,4,6	2010	Oct	Sep	2,4,6	2014	May - Jun	Sep
Egypt GILO	2	2009	Feb	Sep	2	2011	Apr - May	Sep
Georgia GPriEd	2,3,4,5,6	2013	May	Sep	2,3,4,5,6	2015	Feb - Mar	Sep
Ghana PFE Learning	1,2	2017	May - Jun	Sep	1,2	2019	May - Jun	Sep
Haiti TOTAL	1,2	2012	Nov - Dec	Oct	1,2	2013	May - Jun	Oct
Indonesia PRIORITAS (Cohort1)	3	2012	Nov - Dec	Jul	3	2016	Oct - Dec	Jul
Indonesia PRIORITAS (Cohort2)	3	2013	Nov	Jul	3	2016	Oct - Dec	Jul
Kenya EMACK2	1,2,3	2013	Feb	Jan	1,2,3	2014	Sep	Jan
Malawi MEGRA	2,4	2013	May	Sep	2,4	2015	Apr - Jun	Sep
Malawi MTPDS	2,4	2010	Nov	Sep	2,4	2012	Nov	Sep
Nepal EGRP	1,2,3	2016	Feb - May	Apr	1,2,3	2020	Feb - Mar	May
Tanzania TZ21	2	2012	Feb	Jan	2	2014	Sep - Oct	Jan
Yemen YEGRA	1,2	2012	Nov	Sep	1,2	2013	May	Sep
Zambia RTS	2,3	2012	Oct	Jan	2,3	2016	Nov	Jan
Before-and-after								
Afghanistan ACR	2	2017	Apr - Jul	Mar	2	2018	Apr	Mar
Ethiopia TeachII	2	2012	May - Jun	Sep	2	2014	Jun	Sep
Ghana PFE PrePivot	2	2013	Jul	Sep	2	2015	Jul	Sep
Kenya TUSOME	1,2	2015	Jun - Jul	Jan	1,2	2019	Oct	Jan
Lebanon QITABI (Cohort1)	2,3	2015	Nov	Oct	2,3	2018	Apr	Oct
Lebanon QITABI (Cohort2)	2,3	2016	Apr	Oct	2,3	2018	Apr	Oct
Liberia LTTP2	1,2,3	2011	May	Sep	1,2,3	2015	Jun	Sep
Malawi MERIT	1,3	2016	May	Sep	1,3	2018	May - Jun	Sep
Mali PHARE	2,4	2009	Apr - May	Oct	2,4	2011	May	Oct
Mali SIRA	2	2015	May	Oct	2	2018	May	Oct
Mozambique VamosLer	2	2017	Sep	Feb	2	2019	Sep - Oct	Feb
Nigeria NEI+	2,3	2016	May	Sep	2,3	2018	Jul	Sep
Pakistan PRP (Balochistan)	3,5	2013	Oct	Apr	3,5	2017	Sep - Oct	Apr
Pakistan PRP (ICT)	3,5	2013	May	Apr	3,5	2017	Apr - May	Apr
Philippines BASA	2,3	2015	Sep - Oct	Jun	2,3	2017	Feb	Jun
Rwanda L3	1,2,3	2014	Oct	Jan	1,2,3	2016	Oct	Feb
Senegal LPT	1,2	2017	May - Jun	Oct	1,2	2021	May - Jun	Oct
Zambia TTL	2	2012	Oct	Jan	2	2016	Oct	Jan

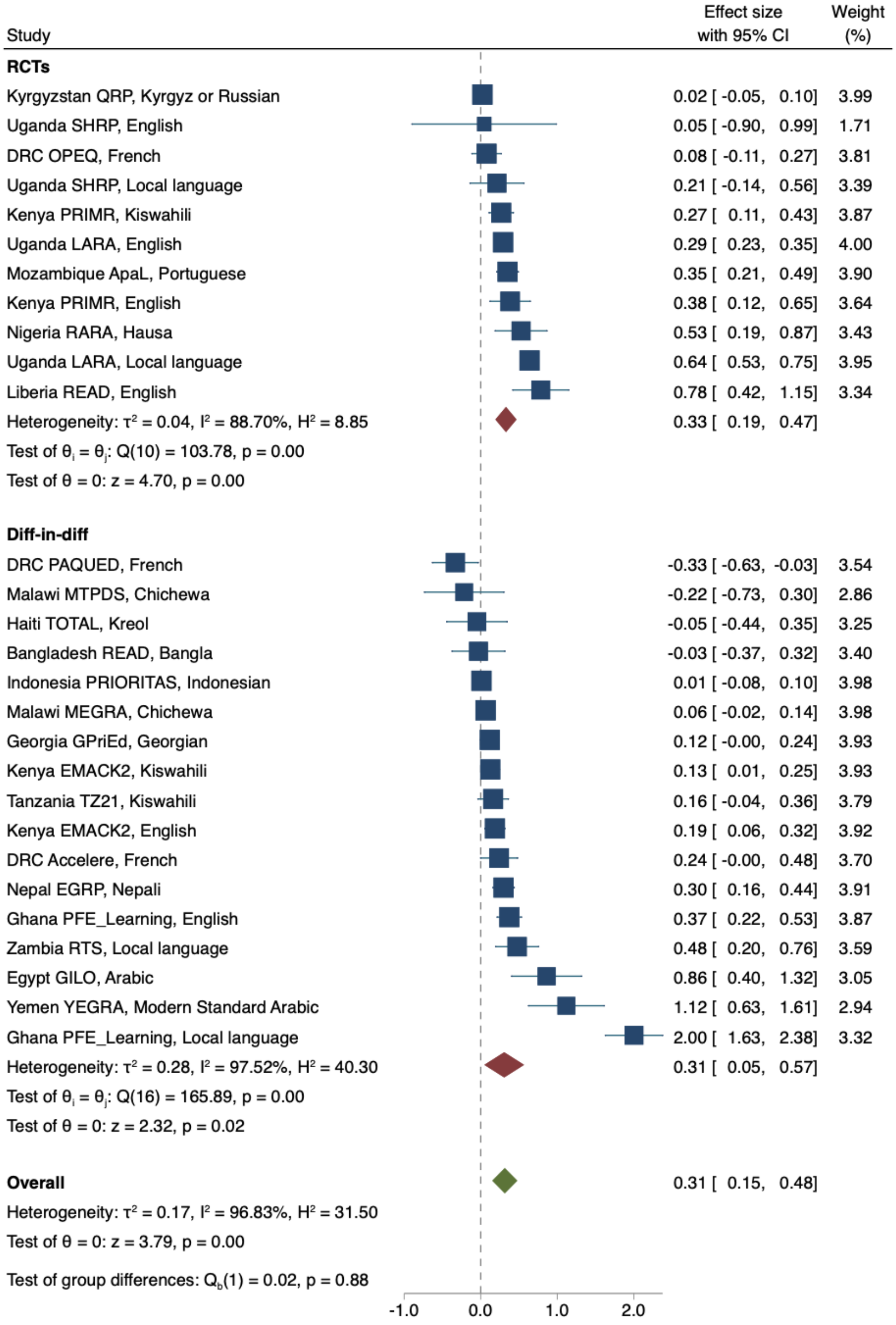
Note: School start month is the month when the specific school year in which the test was conducted started; if it's larger than the Test month, it means that the school year started in that month of the year before the Test month. For example, for the baseline of DRC OPEQ, school started in September,2010 and the test was conducted in March-May,2011. For Pakistan PRP, data from different provinces was collected at different times.

Figure 7: Impacts on oral ready fluency, including implied effects for programs with no control group



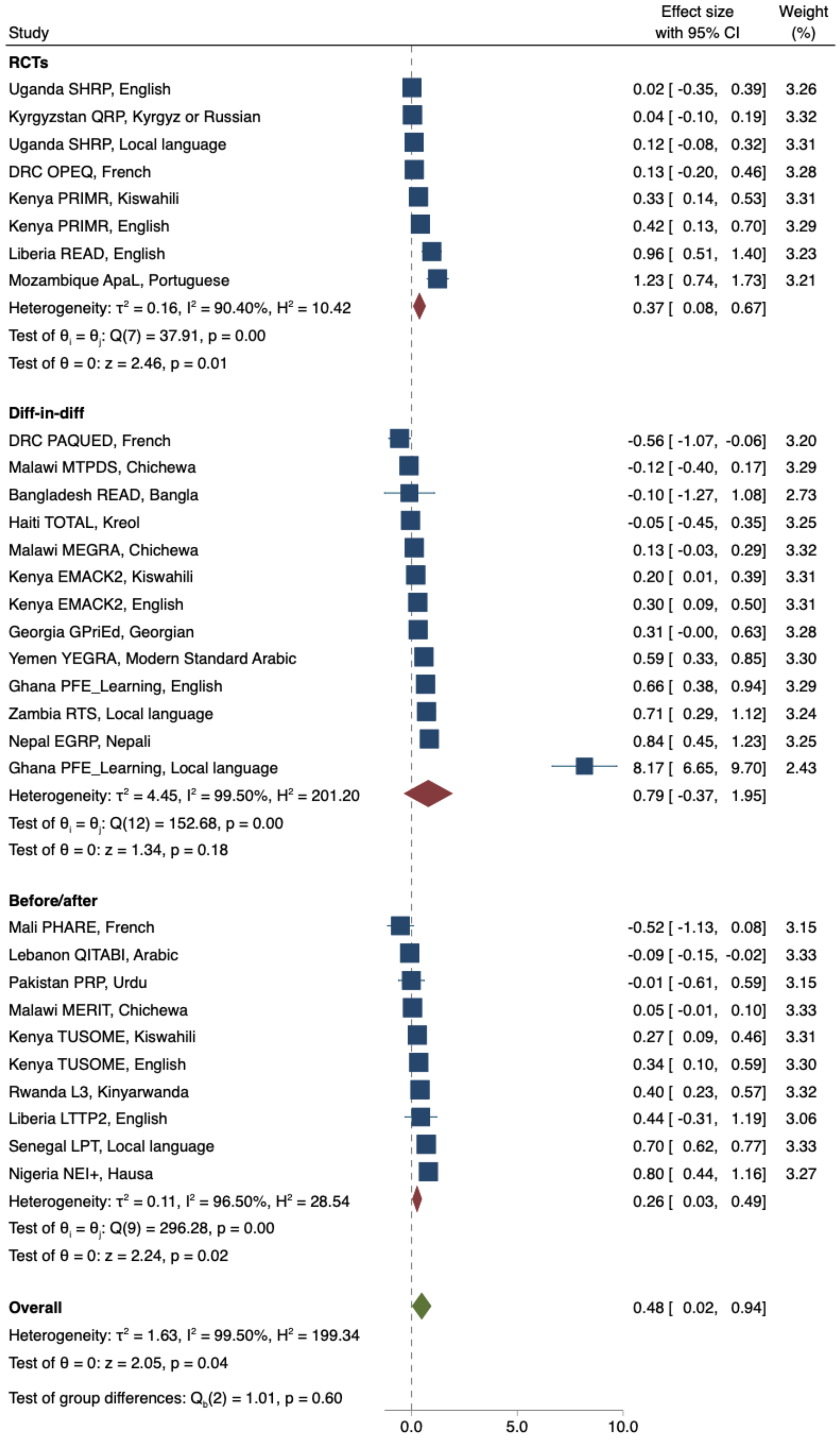
Note: Each coefficient represents an estimated treatment effect, based on the author's reanalysis of microdata. For the before-and-after studies, we report progress in the treatment group after subtracting the mean rate of learning progress in the control groups from experimental and difference-and-difference studies, as described in the text. The outcome for all studies, oral reading fluency, is expressed in correct words per minute (CWPM). Overall effects for each group represent empirical Bayes estimates of a random effects model.

Figure 8: Impacts on oral ready fluency – expressed in standard deviation effect sizes



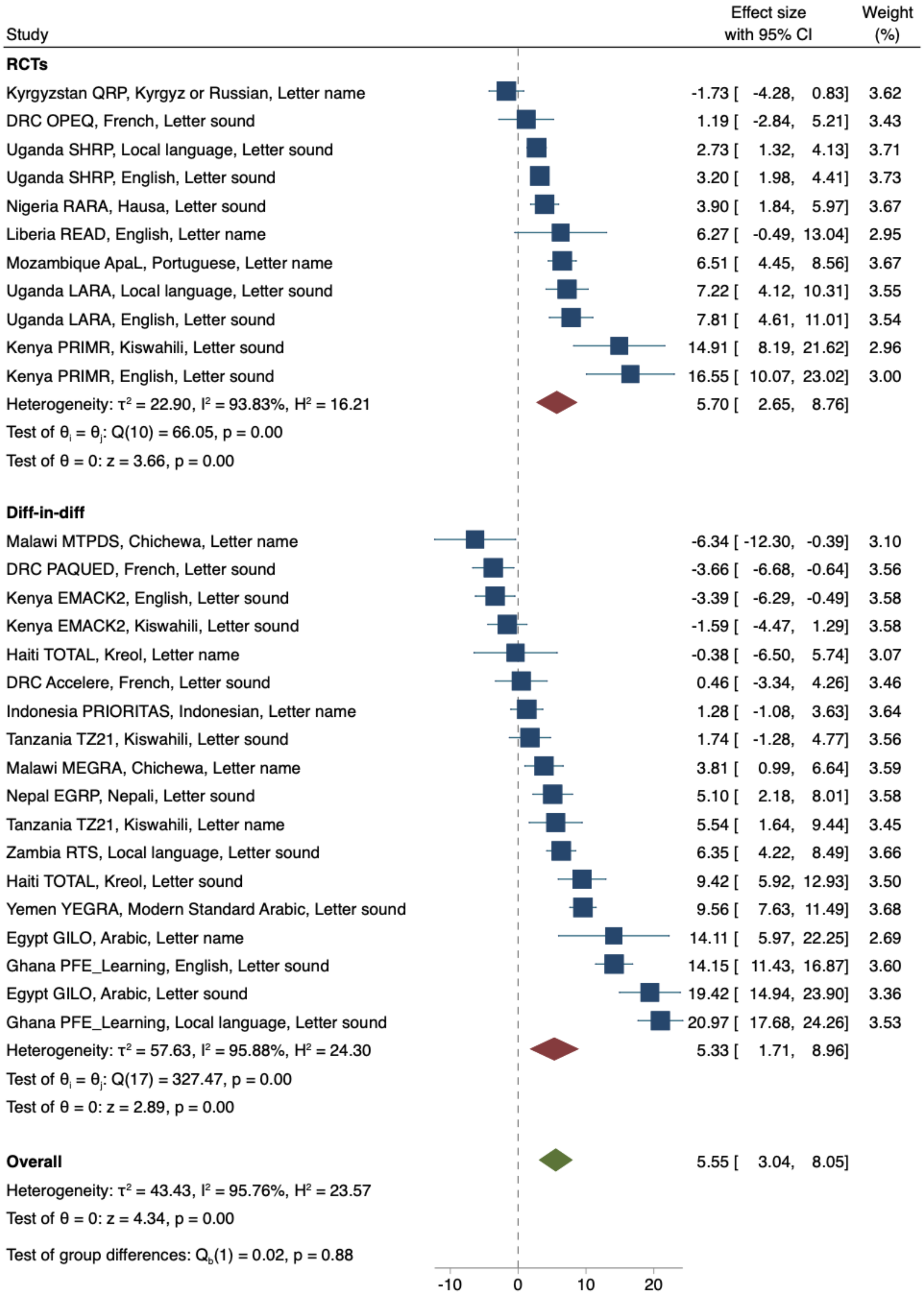
Note: Each coefficient represents an estimated treatment effect, based on the author's reanalysis of microdata. The outcome for all studies, oral reading fluency, is expressed in standard deviations of correct words per minute (CWPM). The standard deviation of CWPM is measured in the control group and/or baseline. Overall effects for each group represent empirical Bayes estimates of a random effects model.

Figure 9: Impacts on oral ready fluency – expressed in ‘equivalent years of schooling’



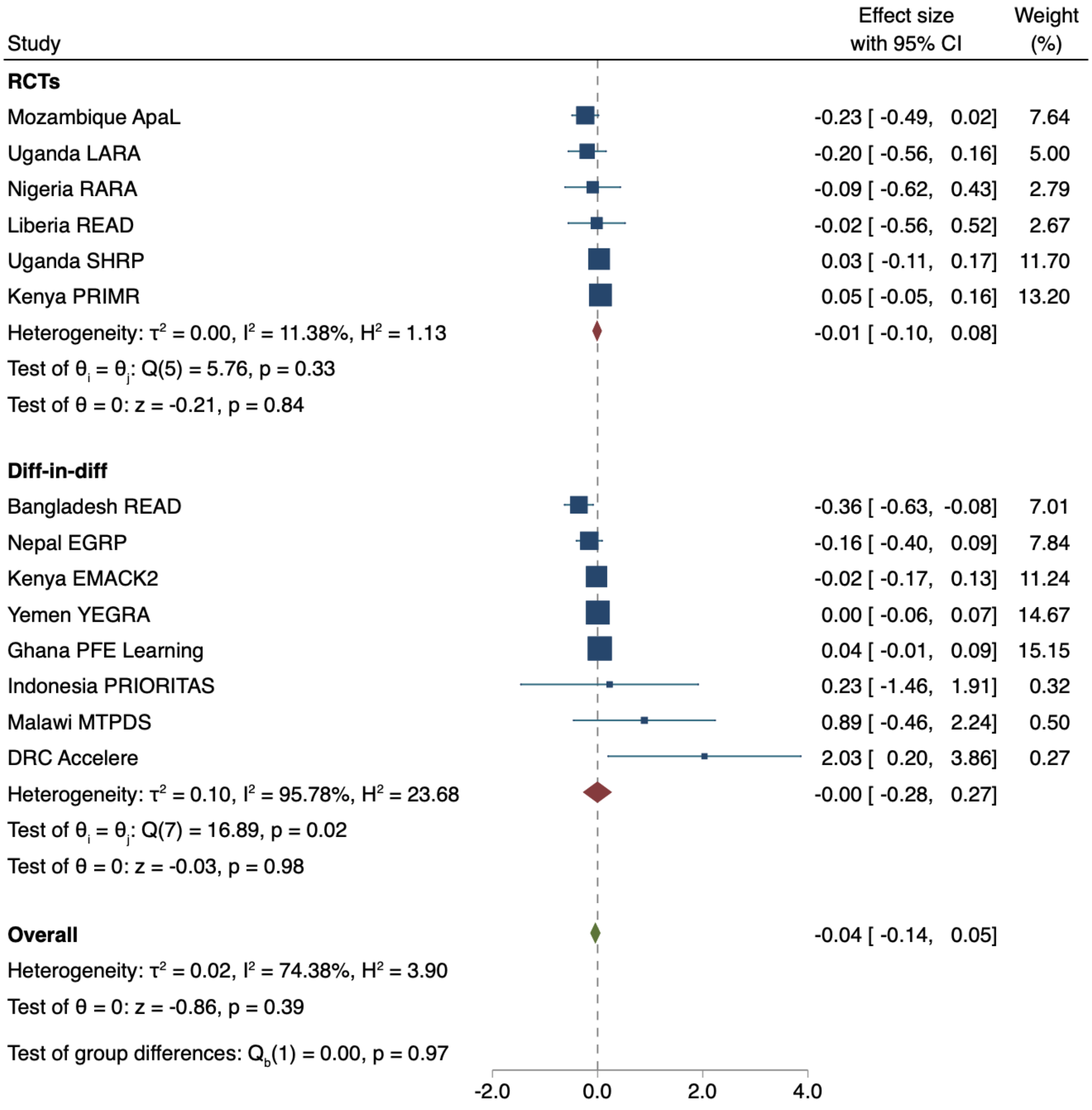
Note: Each coefficient represents an estimated treatment effect, based on the author’s reanalysis of microdata. The outcome for all studies, oral reading fluency, is expressed in equivalent years of schooling (EYOS), i.e., the gain in oral reading fluency divided by the observed difference in oral reading fluence associated with one additional grade of attainment in the control group. (Where no control group is available, the association between grade and oral reading fluency used in the denominator is calculated from the treatment group.) Overall effects for each group represent empirical Bayes estimates of a random effects model.

Figure 10: Impacts on correct letter sounds per minute



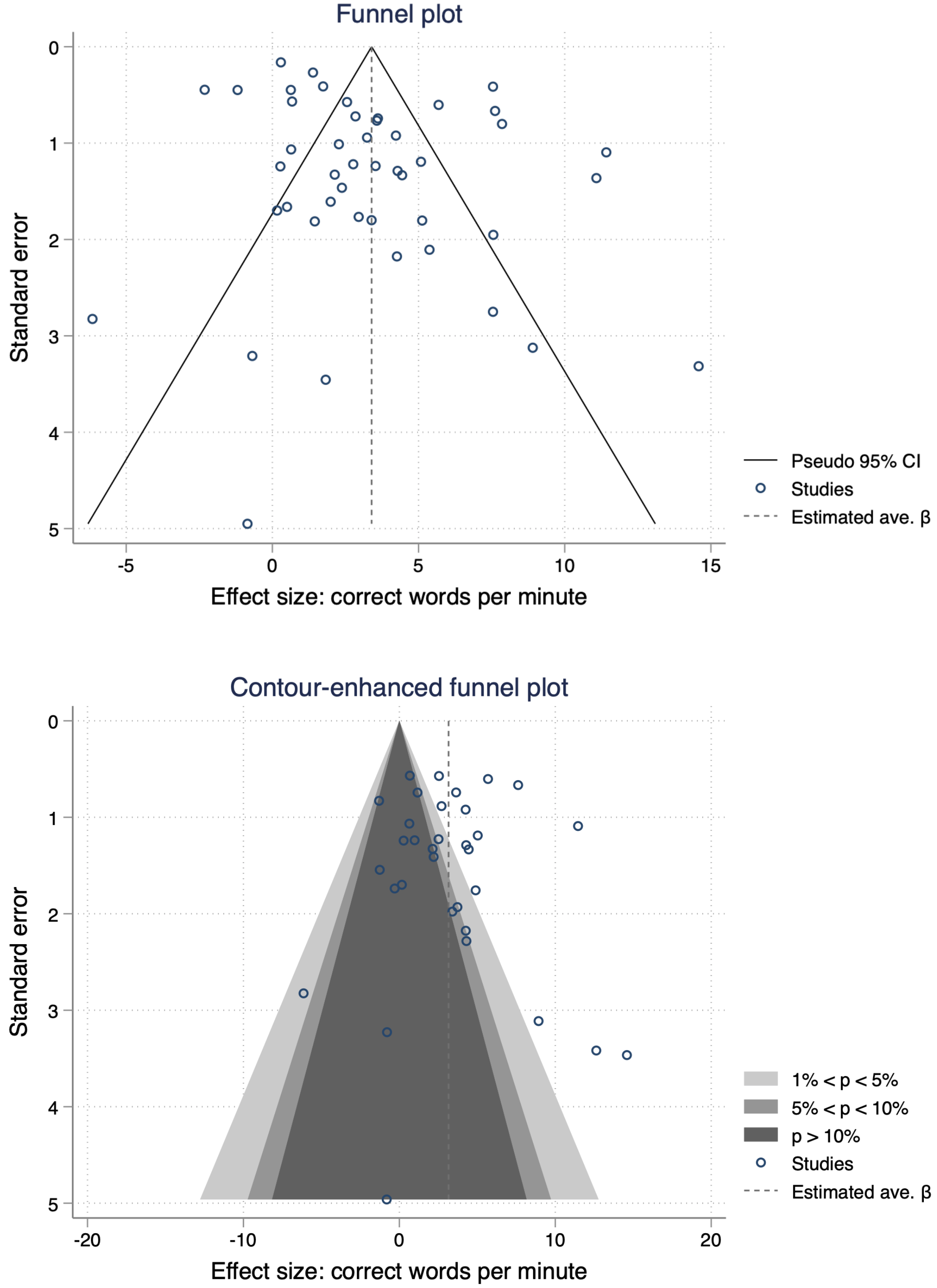
Note: Each coefficient represents an estimated treatment effect, based on the author's reanalysis of microdata. Note that the meta-analysis pools studies using correct letter *sounds* per minute and correct letter *names* per minute. Overall effects for each group represent empirical Bayes estimates of a random effects model.

Figure 11: Test for sorting of pupils in response to treatment: treatment effects on pupil age



Note: Each coefficient represents an estimated treatment effect on pupil age in years. Because endline data collection sampled pupils after treatment (rather than longitudinally tracking pre-treatment samples), in principle treatment could affect average age due to sorting of pupils across schools or grades within a school.

Figure 12: Test for publication bias



Note: The top panel is centered at the overall mean effect size from a random effects model to explore symmetry and 'missing' insignificant or negative effects. The bottom panel is centered at zero to explore p-hacking.