RTI Press

Occasional Paper

ISSN 2378-7996

April 2019

Resetting Targets: Examining Large Effect Sizes and Disappointing Benchmark Progress

Jonathan M. B. Stern and Benjamin Piper

RTI INTERNATIONAL

RTI Press publication OP-0060-1904

RTI International is an independent, nonprofit research organization dedicated to improving the human condition. The RTI Press mission is to disseminate information about RTI research, analytic tools, and technical expertise to a national and international audience. RTI Press publications are peer-reviewed by at least two independent substantive experts and one or more Press editors.

Suggested Citation

Stern, J. M. B., and Piper, B. (2019). *Resetting Targets: Why Large Effect Sizes in Education Development Programs Are Not Resulting in More Students Reading at Benchmark*. RTI Press Publication No. OP-0060-1904. Research Triangle Park, NC: RTI Press. https://doi.org/10.3768/rtipress.2019.op.0060.1904

This publication is part of the RTI Press Research Report series. Occasional Papers are scholarly essays on policy, methods, or other topics relevant to RTI areas of research or technical focus..

RTI International
3040 East Cornwallis Road
PO Box 12194
Research Triangle Park, NC
27709-2194 USA

Tel: +1.919.541.6000
E-mail: rtipress@rti.org
Website: www.rti.org

Cover photo: Alex Kamweru

Contents

About the Authors

Jonathan M. B. Stern, PhD, is a Senior Education Research and Evaluation Specialist at RTI International. His recent work has included the development and validation of reading and mathematics assessments for primary school learners; training and capacity building efforts for ministries of education on school inspection and monitoring, standard setting, and national assessment framework development; and leading research design efforts across projects. Dr. Stern has conducted research and provided technical assistance to programs in more than twenty countries across sub-Saharan Africa, the Middle East, and Asia.

Benjamin Piper, EdD, is the Senior Director of Africa Education for RTI International and is based in Nairobi. He provides support to large-scale education projects across the world. He led a multi-country study of large-scale teacher professional development modalities and is currently leading a study of instructional coaching practices. Dr. Piper was previously the Chief of Party of Tusome, PRIMR, and the National Tablets Programme. He is currently leading a multicountry study of highly effective large-scale education programs with funding from the Gates Foundation.

RTI Press Associate Editor

Meera Viswanathan

Acknowledgments

The authors acknowledge the generous support of an IR&D grant from RTI International that supported the writing of this paper. We would also like to thank everyone who provided feedback and comments on early drafts of this paper. Finally, we are grateful for Erin Newton's thorough editorial contributions.

Abstract

This paper uses recent evidence from international early grade reading programs to provide guidance about how best to create appropriate targets and more effectively identify improved program outcomes. Recent results show that World Bank and US Agency for International Development–funded large-scale international education interventions in low- and middle–income countries tend to produce larger impacts than do interventions in the United States, as measured by effect sizes. However, these effect sizes rarely translate into large gains in mean oral reading fluency scores and are associated with only small increases in the proportion of students meeting country-level reading benchmarks. The limited impact of these low- and middle–income countries' reading programs on the proportion of students meeting reading benchmarks is in large part caused by right-skewed distributions of student reading scores. In other words, modest impacts on the proportion of students meeting benchmarks are caused by low mean scores and large proportions of nonreaders at baseline. It is essential to take these factors into consideration when setting program targets for reading fluency and comprehension. We recommend that program designers in lower-performing countries use baseline assessment data to develop benchmarks based on multiple performance categories that allow for more ambitious targets focused on reducing nonreaders and increasing beginning readers, with more modest targets aimed at improving oral reading fluency scores and increasing the percentage of proficient readers.

Introduction

The goal of nearly all early grade reading interventions is to improve the reading ability of students affected by the program. For US Agency for International Development (USAID)–funded programs in low- and middle–income countries (LMICs), these gains are typically measured by increases in the number and proportion of proficient readers (i.e., learners who demonstrate reading fluency and comprehension of grade-level text). Many programs set benchmarks and targets to ensure that expectations are clear for donors and implementers alike. Benchmarks are defined as performance standards (e.g., an empirically derived number of correct words per minute for oral reading fluency). Targets represent the proportion of students meeting said benchmarks at a prescribed time in the future. Limited guidance is available on how best to set appropriate and achievable targets for improved literacy outcomes, and we fear that inappropriate target setting can have detrimental effects on programs. For this paper, we use recent evidence about reading program outcomes in LMICs to create recommendations for program designers on how to set better targets for program reporting and more effectively identify program outcomes.

Although this paper focuses on measuring the impacts of education programs in LMICs, it is useful to begin by comparing the size of gains in the education development field with those in US education programs, an area with more established research. After describing the relative gains of US and LMIC interventions, we turn to a discussion about the interplay among the considerations needed to fully understand best practices for target setting (i.e., the linkages between effect sizes,[1] mean score gains, and improvements at benchmark, and the impacts of zero scores and skewed performance distributions on increasing the proportion of proficient readers). Finally, we provide conclusions and recommendations for setting targets and measuring program impacts for early grade reading programs in LMICs at both the lower and upper end of the achievement distribution.

[1] Effect sizes provide a standardized measure of the difference in performance between two groups (e.g., treatment and control; intervention and non-intervention). In the simplest sense, an effect size is calculated by subtracting the mean of the non-intervention group from the mean of the intervention group and then dividing by the standard deviation (SD). Accordingly, effect sizes account for variations in the data and in the sample size, and they are reported in SD units. The most commonly cited guideline for interpreting effect sizes asserts that a small effect is greater than 0.2 SD, but less than 0.5 SD; a medium effect is at least 0.5 SD, but less than 0.8 SD; and a large effect is greater than or equal to 0.8 SD (Cohen, 1988). However, these standards are typically higher than those applied to educational interventions (Graham & Kelly, 2018; US Department of Education, 2014).

Evidence from US Education Interventions

We use the extensive literature base that exists on the effectiveness of educational intervention programs in the United States as a starting point for understanding the expected gains of LMIC reading programs. The range of effect sizes across US programs is understandably wide (because of intervention scope, duration, dose, quality, and other related factors, as well as the imprecision of effect size analyses); for example, Slavin, Lake, Chambers, Cheung, & Davis (2009) found a range of -0.53 SD to +1.05 SD across studies, whereas Hattie (2017) found a range of -0.90 SD to +1.57 SD when examining intervention-specific factors. Researchers have conducted meta-analyses to synthesize and average these disparate results across studies (e.g., Camilli, Vargas, Ryan, & Barnett, 2010; Ehri et al., 2001; Hattie, 2009 Fisher, Frey, & Hattie, 2016). One key difference between the US and LMIC research discussed in this paper is the size of the interventions, with recent educational interventions in the LMICs implemented at a scale that far surpassed that of the US interventions studied. Another difference is that the estimated US effect sizes in this paper are based on meta-analyses, whereas those from LMIC interventions are based on individual studies.

Estimates of the average effect sizes for US educational interventions across each level of schooling are presented in Table 1. One comprehensive analysis by Hill, Bloom, Black, & Lipsey (2008) estimated a 0.23 standard deviation (SD) average effect size for lower primary-level education impacts, based on a meta-analysis of 76 other meta-analyses of kindergarten through Grade 12 (K–12) educational interventions. A more limited analysis from a subset of randomized studies within this analysis produced an average effect size of 0.33

 https://doi.org/10.3768/rtipress.2019.op.0060.1904

Table 1. Average effect sizes (in standard deviations) for US education interventions, by level of schooling

Early childhood (national scale)[a]	Lower primary (Grades 1–3)[b]	Primary (randomized only)[b]	Upper primary + secondary reading fluency[c]
0.20 to 0.27	0.23	0.33	0.16 to 0.30

[a] Source: Puma, Bell, Cook, & Heid (2010); Shager et al. (2013).

[b] Source: Hill et al. (2008).

[c] Includes Grades 4 through 12. Source: Scammacca, Roberts, Vaughn, & Stuebing (2015); Wanzek et al. (2013).

Note: We selected the estimates on the basis of their comparability to early grade reading interventions in LMICs in terms of scale, level of schooling, outcome of interest, or a combination of these factors.

SD at the primary school level (Hill et al., 2008). The effect size range of 0.20 SD to 0.27 SD for programs focused on early childhood comes from estimates of the effectiveness of the US Department of Health and Human Services' national Head Start program on a range of cognitive and achievement outcomes (Puma et al., 2010; Shager et al., 2013). Meanwhile, the combined effect size range for upper primary and secondary school of 0.16 SD to 0.30 SD is specific to reading fluency outcomes (Scammacca et al., 2015; Wanzek et al., 2013). We specifically chose these studies because of their relevance and comparability with the trend of early grade reading interventions in education development in LMICs (i.e., moving to scale and focusing on reading fluency). These US domestic results showed a high level of consistency across levels of schooling, with an overall average effect size between approximately 0.20 SD and 0.30 SD.

Ultimately, an effect size of between 0.20 SD and 0.33 SD appears to be the most appropriate range representative of US studies (based on Table 1). Note that these estimates provide an upper-bound (if not inflated) point of comparison for LMIC studies for three reasons. First, we were unable to find many large-scale education interventions at the lower primary or primary level in the United States to meaningfully compare with the many large LMIC programs, which are typically supporting thousands of schools rather than the dozens of schools in most US programs. Second, most effect sizes available for large-scale LMIC interventions stem from non-randomized studies implemented at a large or even a national scale with before and after comparisons rather than RCTs. Third, the US estimates are based primarily on meta-analyses, which are criticized as inflating average effect sizes because of the inclusion of studies that are small, flawed, or both.

When it comes to proficiency, we show later in this paper that children in LMICs struggle to meet the benchmarks set for them. It is worth noting that this is also a problem in the US education system. For example, scores from the 2017 National Assessment of Education Progress (NAEP), known as the Nation's Report Card, showed that only 36 percent of Grade 4 students are reading at NAEP proficiency, which has remained basically unchanged since 2007. Having a relatively small percentage of students meeting a benchmark does not inherently mean that the proficiency standards themselves are inappropriate (as they may represent ideal standards of reading performance). For example, NAEP proficiency is not the equivalent of grade-level performance but is instead based on student competency on challenging subject matter (National Center for Education Statistics, 2018). Therefore, understanding the linkage between the relative difficulty of proficiency standards and the local curriculum and expectations is essential for setting appropriate grade-level benchmarks and challenging, but attainable, targets. In LMICs, it may be that the failure of programs to reach benchmarks is in part due to benchmark levels that are set without adequately considering the country-level characteristics of learning outcomes.

Evidence from International Education Interventions in Low- and Middle-Income Countries

For international primary school interventions in LMICs, Graham & Kelly (2018) calculated an overall estimate of effectiveness from a recent World

 https://doi.org/10.3768/rtipress.2019.op.0060.1904

Bank review of 18 control-group design evaluations for students in Grades 1 through 4. Each of these evaluations used the Early Grade Reading Assessment (EGRA) to measure reading performance. EGRA was developed in 2006 and has since been adapted for use in more than 70 countries and more than 120 languages. The EGRA oral reading subtask was largely informed by the Dynamic Indicators of Basic Early Literacy Skills (RTI International, 2015). Although EGRA measures a range of early literacy skills, the primary focus of the World Bank's analysis was comparing different program's impacts on oral reading fluency. The range and average effect size for impacts on oral reading fluency are displayed in Table 2. The average effect size of 0.44 SD in impacts in LMICs is noticeably larger than the estimated range of impacts from US primary-level interventions. Additionally, according to a new schema set forth by Kraft (2018), most of these LMIC interventions fall into the "easy to scale" category based on their large effect sizes (≥0.20) and low per-student costs (<$500). Notably, the average LMIC effect size is above the 90th percentile in a distribution of 481 effect sizes from education interventions in high-income countries, whereas the average costs fall somewhere below the 20th percentile (Graham & Kelly, 2018; Kraft, 2018; RTI International, 2014).

Table 2. Oral reading fluency effect sizes for international primary education interventions in LMICs (in standard deviations)

	Range	Average
Oral reading fluency effect sizes	0.13 to 0.80	0.44

Source: Graham & Kelly (2018).

In addition to estimating the oral reading fluency effect sizes across these evaluations, Graham & Kelly (2018) calculated average oral reading fluency gains. The results are displayed in Table 3. It is clear from these results that although the evaluations produced effect sizes that were larger than the average US effects (as displayed in Table 1 and Table 2), these impacts were associated with relatively modest causal gains in terms of correct words per minute. More specifically, the average oral reading fluency improvement (treatment over control) in these evaluations was 6.1 correct words per minute (cwpm) over the entire term of the evaluation, with a per year average gain of 4.3 cwpm.

Table 3. Oral reading fluency impact estimates for international education interventions in LMICs (in correct words per minute)

	Range	Average
Oral reading fluency gain overall	0.07 to 21.20	6.1
Oral reading fluency gain per year	0.07 to 14.42	4.3

Source: Graham & Kelly (2018).

Relationship Between Mean Oral Reading Fluency Gains and Effect Sizes

To understand the relationships among effect sizes, oral reading fluency gains, and the ultimate goal of increasing the proportion of fluent readers with comprehension, we examined impacts from six LMIC programs. Although we selected these projects based on the availability of and access to their data, they represent regional diversity (Middle East, sub-Saharan Africa, and South Asia) and a range of implementation scales (pilot, regional, and national). Furthermore, their effect sizes and oral reading fluency gain scores are consistent with the evaluations from the World Bank study. In fact, they are identical, with an average effect size of 0.44 SD and an average oral reading fluency gain of 6.1 cwpm (Table 4). Because national-scale interventions do not have control groups (by design), we conducted all analyses using pre-test to post-test results with treatment groups only). Note that all results are from Grade 2 analyses.

An essential point to note from Table 4 is that larger Grade 2 effect sizes are associated with higher oral reading fluency gain scores (with a correlation of 0.87), as one would expect. What is important to understand, however, is that the relationship between oral reading fluency effect sizes and mean oral reading fluency gains can differ depending on the distribution of reading abilities in a given country and the baseline performance level of a country or program. For example, Kenya (Freudenberger & Davis, 2017) and Malawi had nearly identical effect

 https://doi.org/10.3768/rtipress.2019.op.0060.1904

Table 4. Selected country impact data: Oral reading fluency effect sizes and mean gains

Country	Scale	Language	Oral reading fluency effect size (standard deviations)	Oral reading fluency mean gain (correct words per minutes)
Jordan	Pilot	Arabic	0.49	5.8
Kenya[a]	National	Kiswahili	0.71	11.0
Malawi	Pilot	Chichewa	0.70	7.0
Nepal	National	Nepali	0.20	3.0
Rwanda[b]	National	Kinyarwanda	0.30	5.7
Uganda	Regional	Luganda	0.23	4.1
Average			**0.44**	**6.1**

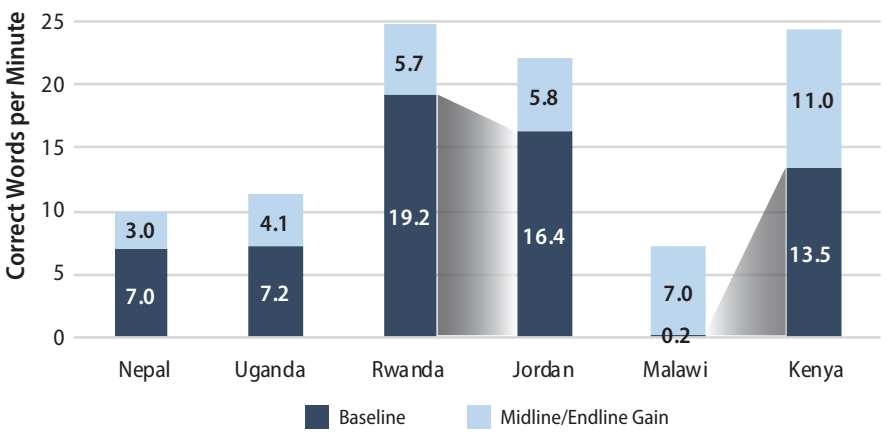
[a] Freudenberger & Davis (2017).

[b] Education Development Center (2017).

Note: Unless otherwise noted, we calculated estimates directly from RTI International project data. All analyses focused on local or national languages; international languages were not included in these analyses. Jordan and Malawi will eventually have large-scale impact evaluation data released that can replace this pilot data.

sizes, but the oral reading fluency gain was 11.0 cwpm for Kenya, whereas it was only 7.0 cwpm for Malawi. Conversely, the oral reading fluency gains were similar for Jordan (5.8 cwpm) and Rwanda (5.7 cwpm) (Education Development Center, 2017), but these impacts led to very different effect sizes (Jordan with 0.49 SD versus Rwanda with 0.30 SD).

Figure 1. Within-project gains identified at midline or endline above baseline: Grade 2 oral reading fluency rate, by country



Note: Shading points to differences in baseline values between Jordan and Rwanda and between Malawi and Kenya.

These apparent inconsistencies can be explained, in part, by Figure 1. Kenya had a much higher baseline value than Malawi (13.5 cwpm vs. 0.2 cwpm) and more variation in the data. Malawi's lower baseline level (combined with the lack of variation in scores) means that a smaller increase in mean oral reading fluency was needed to produce the same effect size as Kenya. Similarly, Rwanda's higher starting point (19.2 cwpm) appears to have led to a smaller effect size than Jordan, even though both locations showed approximately the same oral reading fluency gain (5.7 or 5.8 cwpm).

.Ultimately, these surprising findings can be traced to the fact that lower mean baseline values in these contexts often occur as a result of right-skewed distributions, larger proportions of nonreaders (i.e., floor effects), and therefore less variation in the data. Recall that effect sizes are calculated by determining the difference between two groups and dividing that difference by the standard deviation. As an example, if two countries produced the same mean improvement (e.g., 10 cwpm), but Country A had twice the variation as Country B, then the effect size for Country B would be twice as large (i.e., $10/x$ would be two times as large as $10/2x$). This difference has clear implications for how program designers should set targets and estimate expected program impacts.

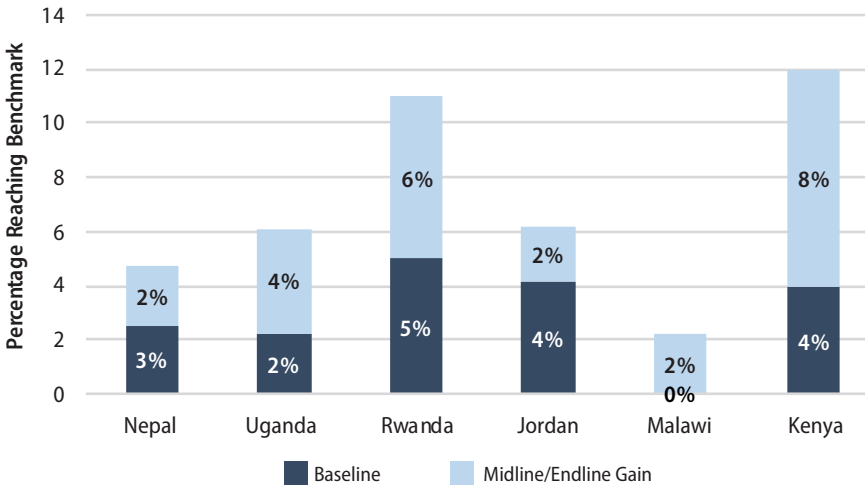
 https://doi.org/10.3768/rtipress.2019.op.0060.1904

Relationships Between Mean Scores, Effect Sizes, and Proficiency Benchmarks

Although nearly all early grade reading interventions aim to improve mean oral reading fluency scores, using mean gains as the key measure can produce inappropriate or misleading measures of program performance when working with skewed performance distributions. Furthermore, the goal of these programs typically is to produce as many grade-level readers as possible (i.e., those who can read text that is appropriate for their age and grade, with fluency and comprehension). Setting an objective of increasing the number or proportion of students reading at grade level has the added benefit of alignment with the United Nations' Sustainable Development Goal 4.1.1(a). This goal calls on countries to measure and report on the percentage of children achieving at least minimum proficiency in reading and mathematics at the end of Grade 2 or 3.

To estimate program impact on reaching grade-level targets, we examined the change in the proportion of Grade 2 students reading at or above the fluency benchmark for each country and language. Although all benchmarks should be based on language-specific evidence of what constitutes a proficient reader, for the purposes of our analysis, we used the range of the programs with existing benchmarks to "impute" benchmarks for the two countries without established benchmarks (Nepal and Uganda).[2] The benchmarks were as follows: Malawi, 40 cwpm; Nepal, Uganda, and Kenya, 45 cwpm; Jordan, 46 cwpm; and Rwanda, 47 cwpm.

Figure 2. Grade 2 percentages meeting benchmark over time, by country



Program impact on the percentage of children reading at benchmark ranged from a low of 2 percentage points in Nepal, Jordan, and Malawi, to a high of 8 percentage points in Kenya. The average increase was 4 percentage points in the proportion of students reading at benchmark across all countries (Figure 2). Jordan and Malawi had two of the largest increases in mean scores and effect sizes but were tied for the smallest increases in the percentage of children reading at benchmark. The most striking situation was Malawi, where a 7.0 cwpm increase (and an effect size of 0.70 SD) led to an increase of only 2 percentage points in the proportion of students reaching the 40 cwpm benchmark. These mismatched results were due almost entirely to the large proportion of students with zero scores at baseline, and the skewed distribution of scores. Given this skew, Malawi had a very small proportion of students near the benchmark (or within "striking distance") at baseline. As a result, the relatively large program effect size and significant oral reading fluency gains ultimately had little impact on benchmark performance.

Table 5 presents the percentage of students with a zero score on oral reading fluency at baseline and the reduction in zero scores over the life of the evaluation alongside some of the values from Table 4 and Figure 2. Zero scores (and nonreaders) are calculated strictly as students who are unable to read a single word correctly in the oral reading passage. In every country, the percentage point

[2] The imputed benchmarks for Nepal and Uganda in this paper are solely for illustrative purposes, and only affect the analyses focused on improvements at benchmark. To ensure that these findings were not simply an artifact of where the imputed benchmarks were set, we conducted sensitivity analyses by varying the imputed benchmarks from a low of 35 cwpm to a high of 90 cwpm. Decreasing the imputed benchmarks produced no significant changes for either country, but increasing the benchmarks further reduced the gains for both countries (thus strengthening our argument regarding the disconnect between effect sizes, mean oral reading fluency gains, and benchmark performance for proficient readers).

https://doi.org/10.3768/rtipress.2019.op.0060.1904

Table 5. Selected country Grade 2 impact data for all oral reading fluency estimates

Country	Oral reading fluency score gain (correct words per minute)	Percentage point increase in students reaching oral reading fluency benchmark	Percent of oral reading fluency zero scores at baseline	Percentage point reduction in oral reading fluency zero scores	Effect size (standard deviations)
Jordan	5.8	2	17%	10	0.49
Kenya	11.0	8	43%	24	0.71
Malawi	7.0	2	98%	34	0.70
Nepal	3.0	2	62%	14	0.20
Rwanda	5.7	6	33%	7	0.30
Uganda	4.1	4	64%	10	0.23
Average	**6.1**	**4**	**53%**	**16**	**0.44**

reduction in zero scores was larger than the percentage point increase in students reaching the oral reading fluency benchmark. In most cases, the difference between percentage point reductions in zero scores and percentage point increases in students reaching the benchmark was quite large (particularly for Malawi, with a 34 percentage point reduction in nonreaders and a 2 percentage point increase in fluent readers). On average, although the countries were able to increase the proportion of fluent readers by only 4 percentage points, they were able to reduce the proportion of nonreaders by a substantial 16 percentage points. This finding shows that significant progress was being made across these interventions, despite limited improvements in the proportion of students reading fluently. This finding also provides evidence of the value of using multiple performance thresholds in the benchmarking process to account for improvements at varying levels of performance (e.g., nonreaders, beginners, proficient, advanced). This approach aligns with international best practices for benchmark or standard setting (Bejar, 2008; Cizek, 1996; Stern, Dubeck, & Dick, 2018).

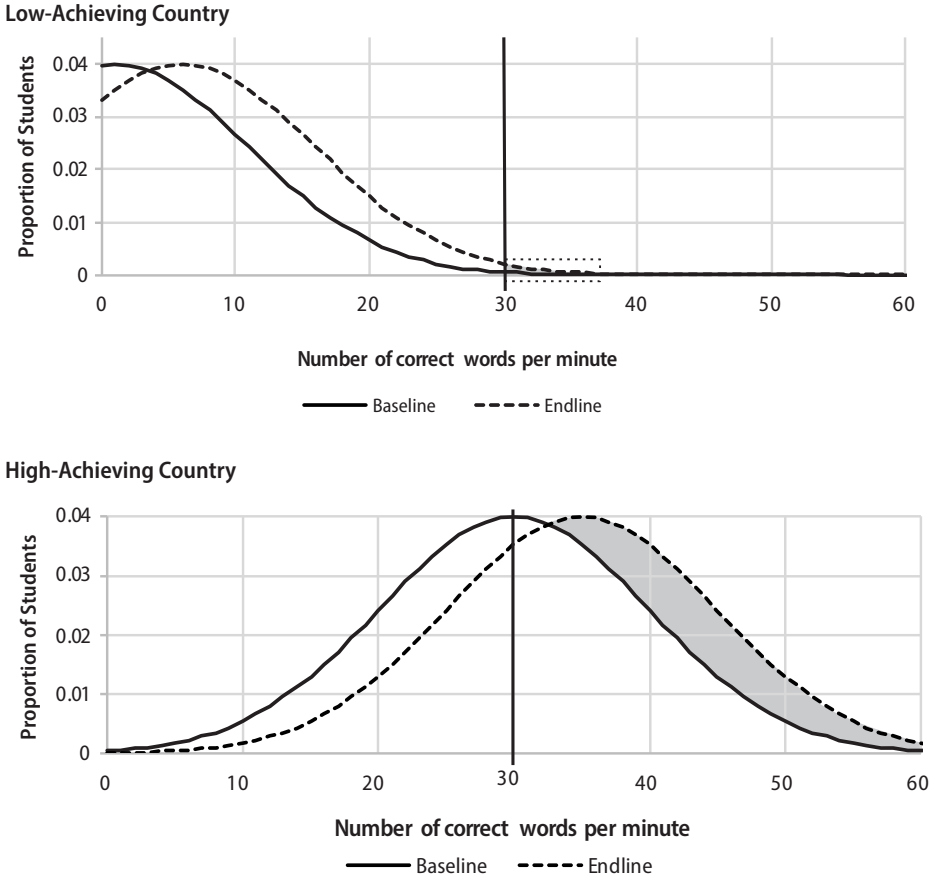
Impact of Zero Scores on Proficient Reader Benchmark Gains

Much of the explanation for why relatively large effect sizes tend to result in relatively small increases in the proportion of students reading fluently stems from the difficulty in producing shifts in the distribution of student scores, given the baseline status of that distribution. An example of the relationship between baseline student achievement level and the sensitivity of benchmarks is illustrated in Figure 3, which shows hypothetical distributions of oral reading fluency in low-achieving and high-achieving countries. In both countries, the proficient reader fluency benchmark was set at 30 cwpm, and there was a mean improvement of 5 cwpm from baseline to endline. In the low-achieving country, this improvement led to an increase of only 0.006 percentage points (i.e., from 0.002 percent to 0.008 percent) in the students meeting the benchmark. This improvement can be seen by the small area between the two curves that occurs above the benchmark (see small shaded area between the curves to the right of the vertical line in Figure 3). In contrast, in the high-achieving country, the same mean improvement in oral reading fluency led to an increase of 19 percentage points in the students meeting the benchmark, from 50 percent to 69 percent (with a significantly larger area between the curves above the benchmark). Because so few students were approaching the benchmark at baseline in the low-achieving country example, the entire distribution would require a large shift to the right to produce a significant proportion of new "readers." By contrast, only a small shift in the distribution would be required when many students were already close to the benchmark, as shown in the high-achieving country example.

In many low-income countries, the situation is more similar to that of the hypothetical low-achieving sample in Figure 3. For example, the solid line in Figure 4 represents the distribution of student

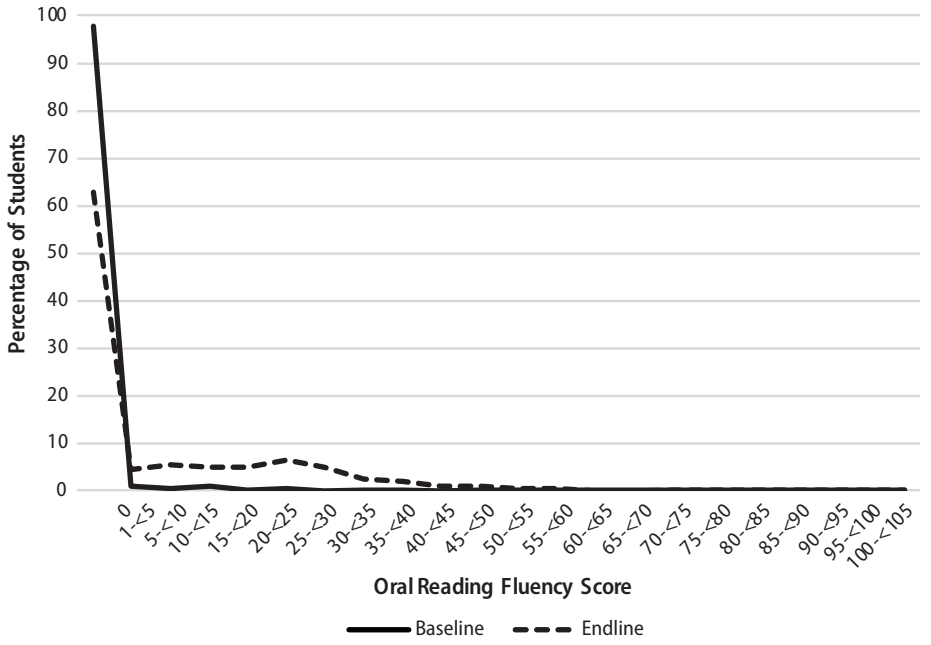
oral reading fluency scores as measured at baseline in Malawi. Because 98 percent of students were unable to read even a single word correctly in one minute, there were few remaining students at any point in the distribution (thus creating basically a flat line at the 0 percent mark for nearly all other score categories). By endline, the proportion of nonreaders significantly decreased to 64 percent. However, it is clear from the dashed line (i.e., endline distribution) that nearly the entire shift in scores took place below the 40 cwpm mark. In other words, although more students began to score between 1 and 40 correct words per minute from baseline to endline, the increase above 40 cwpm was minimal (only 2 percentage points). As with the hypothetical example above, the proportion of students reading at benchmark would increase only by moving students from the very left end of the distribution all the way to the benchmark and above. Meanwhile, the proportion of students reading 10 or more correct words per minute increased from less than 1 percent at baseline to 27 percent at endline (because it is possible to have much larger shifts in the distribution closer to where the majority of students score at baseline). This phenomenon is true in other contexts where the proportion of zero scores is high at baseline, as well as in

Figure 3. An example of students reaching fluency benchmarks in low- and high-achieving countries (hypothetical)



Note: The shaded area between the curves represents the increased proportion of students at benchmark.

Figure 4. Distribution of Grade 2 oral reading fluency scores in Malawi (baseline and endline)



 https://doi.org/10.3768/rtipress.2019.op.0060.1904

distributions with the vast majority of students who were nonreaders or slow readers.

In short, the evidence indicates that perhaps the focus of programs in countries with lower baseline scores should be on reducing the proportion of nonreaders and increasing the proportions of beginning and intermediate readers, until obtaining the benchmark is more realistically within reach. This does not mean, however, that countries should reduce empirically based standards and benchmarks simply because few students are reading at proficient levels. A multiple benchmark approach could help programs to identify meaningful gains in literacy skills when children score very poorly at the baseline.

Conclusions

Although recent evidence has shown that reading interventions in low-income and LMICs tend to produce much larger average effect sizes than US interventions do, it is essential for donors, development partners, and policy makers to understand how these effect sizes translate into contextually relevant findings. For example, these larger effect sizes are associated with modest gains in mean oral reading fluency, with recent estimates averaging approximately 6 cwpm over the life of an intervention evaluation cycle. Furthermore, these oral reading fluency gains are associated with small increases in the proportion of students meeting proficiency-based reading standards or benchmarks (i.e., reading fluently with comprehension), which is the goal of most early grade reading interventions and is an essential outcome indicator for most funding agencies.

The relatively small effect on increased readers is almost entirely due to a combination of low mean scores, minimal variation, and large proportions of nonreaders at baseline (i.e., floor effects). For example, if more than 90 percent of students are nonreaders at baseline (as was the case in Malawi), and the remaining students were nearly all scoring significantly below a given reading fluency standard (e.g., scoring between 0 and 15 cwpm), then even a relatively large increase in the average oral reading fluency score (e.g., 7 cwpm) would not be enough to move a substantial number of students beyond a benchmark of 40 cwpm. In the case of Malawi, this improvement in the average oral reading fluency score led to an increase of only 2 percentage points in the proportion of students at benchmark. However, the 7 cwpm increase was obtained by significantly reducing zero scores (from 98 percent to 64 percent) and producing a larger proportion of beginning readers. Countries such as Indonesia and the Philippines that have higher baseline achievement levels will have different distributions of learning outcomes. Though these countries may experience larger changes in the percentage of children at the benchmark with relatively small mean gains, additional research is required to determine the growth trajectories of children in those countries for setting appropriate targets.

Ultimately, it is essential for education program planners and designers to take the following factors into consideration when setting targets for program impacts: (1) the distribution of scores; (2) the proportion of students at benchmark; (3) the proportion of students at zero (i.e., nonreaders); and (4) the mean fluency score. They could, for example, set ambitious targets for effect sizes, reductions in the proportion of nonreaders, and increases in the proportion of beginning and intermediate readers. They could also establish only modest targets for improvements in mean oral reading fluency scores. Education program planners, designers, and implementers should also consider developing multiple threshold benchmarks so progress can be measured at different points in the distribution, thereby allowing for more nuanced examinations of gains. This multiple threshold approach would follow the practices used by most large-scale international assessments, such as the Trends in International Mathematics and Science Study, the Progress in International Reading Literacy Study, the Programme for International Student Assessment, and the NAEP—all of which report findings across multiple categories.

Finally, program planners should be particularly careful not to overestimate the potential impacts of a program based on the proportion of students reading fluently, particularly in contexts where significant

portions of the student population are classified as nonreaders. The only exception would be if the distribution of student performance clearly showed a reasonable proportion of students at or approaching the benchmark. It is important to note that these considerations are all related to setting and adjusting targets. We do not recommend that countries lower their standards or benchmarks to ensure that greater proportions of students reach them.

Implementers, donors, program planners, and government education leaders, along with teachers, parents, and societies as a whole, share the common objective of putting successful programs in place to ensure that as many students as possible will become proficient readers. However, unrealistic expectations and unachievable targets can only be detrimental to the education sector. When programs overpromise and therefore underdeliver, stakeholders, including funders, naturally come to regard them as not worth the ongoing investment. For this reason, program planners should carefully review the implications of program-based targets and the selection of indicators. They should consider baseline levels of performance and focus lower-performing countries on reducing the percentages of nonreaders and focus higher-achieving countries on improved oral reading fluency. To be clear, this does not mean that countries should lower their aspirational goals of ensuring that all children reach proficiency across early grade reading skills. Evidence-based intervention programs are the means to achieve those goals, but their success must be measured against incremental milestones and targets that are realistic and achievable.

References

Bejar, I. I. (2008). Standard setting: What is it? Why is it important. *R&D Connections, 7*, 1–6.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record*, *112*(3), 579–620.

Cizek, G. J. (1996). Standard-setting guidelines. *Educational Measurement: Issues and Practice*, *15*(1), 13–21. https://doi.org/10.1111/j.1745-3992.1996.tb00802.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Education Development Center (EDC). (2017, January). *Literacy, Language and Learning Initiative (L3)—National fluency and mathematics assessment of Rwandan schools: Endline report*. Prepared for US Agency for International Development/Rwanda under Contract No. AID-696-A-11–00006. Washington, DC: Author. Retrieved from https://pdf.usaid.gov/pdf_docs/PA00MJB7.pdf

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*(3), 250–287. https://doi.org/10.1598/RRQ.36.3.2

Fisher, D., Frey, N., & Hattie, J. (2016). *Visible learning for literacy, grades K-12: Implementing the practices that work best to accelerate student learning*. Thousand Oaks, CA: Corwin Press.

Freudenberger, E., & Davis, J. (2017). *Tusome external evaluation—midline report*. Prepared for the Ministry of Education of Kenya, US Agency for International Development/Kenya, and the UK Department for International Development under Contract No. AID-615-TO-16–00012. Washington, DC: Management Sciences International (MSI), a Tetra Tech company. Retrieved from https://pdf.usaid.gov/pdf_docs/PA00MS6J.pdf

Graham, J., & Kelly, S. (2018). *How effective are early grade reading interventions? A review of the evidence*. Washington, DC: World Bank. Retrieved from. https://doi.org/10.1596/1813-9450-8292

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York, NY: Routledge Press.

Hattie, J. (2017). *Visible learning: 250+ influences on student achievement*. Retrieved from https://visible-learning.org/wp-content/uploads/2018/03/VLPLUS-252-Influences-Hattie-ranking-DEC-2017.pdf

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

RTI International. (2014). *Costing early grade reading programs: An examination of various costs and issues around costing*. Washington, DC: US Agency for International Development.

RTI International. (2015). *Early Grade Reading Assessment (EGRA) toolkit* (2nd ed.). Washington, DC: US Agency for International Development. Retrieved from https://www.globalreadingnetwork.net/sites/default/files/resource_files/EGRA%20Toolkit%20Second%20Edition.pdf

Kraft, M. A. (2018). *Interpreting effect sizes of education interventions*. Brown University working paper. Retrieved from https://scholar.harvard.edu/files/mkraft/files/kraft_2018_interpreting_effect_sizes.pdf

National Center for Education Statistics. (2018). *2017 reading results*. Retrieved from https://www.nationsreportcard.gov/reading_math_2017_highlights/files/infographic_2018_reading.pdf

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start impact study (final report)*. Prepared for the Office of Planning, Research and Evaluation, Administration for Children and Families, US Department of Health and Human Services, under Contract 282–00–0022. Rockville, MD: Westat and others. Retrieved from https://www.acf.hhs.gov/opre/resource/head-start-impact-study-final-report

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369–390. https://doi.org/10.1177/0022219413504995

Shager, H., Schindler, H., Magnuson, K., Duncan, G., Yoshikawa, H., & Hart, C. (2013). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*, *35*(1), 76–95. https://doi.org/10.3102/0162373712462453

Slavin, R. E., Lake, C., Chambers, B., Cheung, A., & Davis, S. (2009). *Effective beginning reading programs: A best evidence synthesis*. Prepared for the Institute of Education Sciences, US Department of Education, under Grant No. R305A040082. Baltimore, MD: Center for Data-Driven Reform in Education, Johns Hopkins University. Retrieved from http://www.bestevidence.org/word/begin_read_jan_26_2009.pdf

Stern, J. M. B., Dubeck, M. M., & Dick, A. (2018). Using Early Grade Reading Assessment (EGRA) data for targeted instructional support: Learning profiles and instructional needs in Indonesia. *International Journal of Educational Development*, *61*, 64–71. https://doi.org/10.1016/j.ijedudev.2017.12.003

US Department of Education. (2014). *What Works Clearinghouse: Procedures and standards handbook version 3.0*. Washington, DC: Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf

Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, *83*(2), 163–195. https://doi.org/10.3102/0034654313477212

RTI International is an independent, nonprofit research institute dedicated to improving the human condition. We combine scientific rigor and technical expertise in social and laboratory sciences, engineering, and international development to deliver solutions to the critical needs of clients worldwide.

www.rti.org/rtipress **RTI Press publication OP-0060-1904**