EDDATA II

Improvements in Reading Skills in Kenya: An Experiment in the Malindi District

EdData II Technical and Managerial Assistance, Task Number 4
Contract Number EHC-E-04-04-00004-00

Strategic Objective 3
March 2009

This publication was produced for review by the United States Agency for International Development. It was prepared by RTI International in collaboration with the Aga Khan Foundation.

Improvements in Reading Skills in Kenya: An Experiment in the Malindi District

Prepared for
Bureau for Economic Growth, Agriculture and Trade (EGAT/ED)
United States Agency for International Development

Prepared by
Luis Crouch and Medina Korda
RTI International
3040 Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194

and

David Mumo
Aga Khan Foundation

Table of Contents

Page

List of Figures

Page

List of Tables

Page

Acknowledgments

Most funding for both the intervention and the tracking for this effort was generously provided by USAID/Kenya and USAID/Washington. The authors and their colleagues at RTI and Aga Khan Foundation, and other associates such as Sylvia Linan-Thompson, also devoted considerable personal "sweat equity" to the activities described herein. The authors are grateful for all the comments. All errors and omissions remain theirs. The opinions expressed in this paper represent only the authors' points of view and are not necessarily to be associated with their home institutions or the funder. For further information (data, instruments, baseline, etc.) please contact the authors: lcrouch@rti.org, mkorda@rti.org, dmumo@yahoo.com.

1. Summary

This paper describes an experimental reading improvement trial in the Malindi District in Kenya, carried out by the Aga Khan Foundation (AKF) and RTI International. The project was funded by the U.S. Agency for International Development (USAID) under the Education for Marginalized Children in Kenya (EMACK II) and EdData II projects and used a randomized assignment of schools into treatment and control schools, with both pre- and post-treatment measurements relying on the Early Grade Reading Assessment (EGRA) tool as an assessment vehicle.[1] After less than a year of reading interventions, rather large improvements in some reading scores—as much as 80%—were noted.[2]

Several considerations, each discussed in detail below, suggest attenuating this increase somewhat, including the fact that the post-treatment measure was performed some two months further into the school calendar than the pre-treatment measure. A far more interesting phenomenon was uncovered, however, with strong substantive and evaluation methodology implications. That is, while the increases found were large, they were found to happen in both the control and treatment schools, and in certain schools much more than in others, in both control and treatment groups.

This naturally puzzled the researchers, who called for further qualitative or "forensic" research. This further qualitative research showed that, unlike in medical trials where a placebo (colloquially, the "sugar pill") allows for a "clean" control group, in Malindi some of the schools not receiving the treatment saw the effects of the treatment on the treatment group, and managed, one way or another, to begin treating themselves, as it were.[3] Furthermore, improvement was very marked in some schools, and the improvement could quite easily be tracked, in a "forensic" causal sense, to the spread of the innovation. This strongly suggests that the impact was not a general drift caused by some third factor.

Taking these various factors into account, the research thus strongly suggests that not only was the treatment impact real, but that the treatment was practical enough, and its impacts observable enough, that teachers in control essentially demanded, and managed to get, treatment as well (although in some cases the teachers deemed it too difficult to bother with). The suggestion that, for methodological reasons, in subsequent trials the control schools be placed further away from the treatment

[1] See http://www.equip123 net/webarticles/anmviewer.asp?a=348 or http://pdf.usaid.gov/pdf_docs/PDACI056.pdf (accessed March 13, 2009) for a description of the EMACK II project. See http://www.eddataglobal.org/documents/index.cfm?fuseaction=searchCountry (accessed March 13, 2009) for a description of the EdData II project. See the same website, EGRA link, for a description of the EGRA tool.

[2] In this paper, all differences whose significance is remarked upon from a substantive point of view are statistically significant. Statistical significance is not remarked upon separately from substantive significance. If results are not statistically significant, they are generally simply not noted at all.

[3] The desire to implement EGRA in some of the control schools was almost of a "subversive" nature (in the positive sense of "subversive"), with at least one "control" principal threatening to organize his own workshop on the methods, using the treatment teachers, if he did not get information about what the treatment schools were doing.

schools has some appeal, but this use of randomization actually militates against the whole idea of randomization, namely that the treatment and control schools be as similar as possible to each other, aside from the fact of the treatment.

The substantive implication is that an approach to reading that is direct, simple, explicit, and intense (or at any rate more intense than what appears to have been common), seems to be welcomed by teachers, principals, and parents, and appears quite capable of boosting performance, on some measures, quite quickly. It is also possible that some of the effect on the control schools was an accountability effect, as will be discussed below. Further research could discover how much of the improvement is due to a pure accountability effect. RTI is involved in such research in Liberia, where one of the treatment groups (a "light" treatment group) is subject only to repeated measurement, is told that there will be repeated measurement, and must report to the community on the results of the measurement. A "full treatment" set of targeted schools in Liberia receives the measurement and, in addition, a treatment similar to—but more intense than—what was tried in Malindi.

In short, the results are quite interesting and potentially very meaningful. Reading results in Malindi improved in about one year—although children are still below a proper benchmark. The contamination or leakage of technique into the control group muddies the results a bit, but further qualitative evidence actually unearthed conclusions that to us seem even more interesting than that "the approach seems to work," because this leakage, and the research into why and how it happened, tells us a good bit about why and how innovations spread, at least under similar conditions.

2. Background and Purpose

Multilateral and bilateral agencies working on education face increasing pressure to focus not just on access (enrollment and completion), but also on quality. The Millennium Development Goals (MDGs) and Education For All (EFA) goals tend to refer to quality, but use no indicators to track quality of education, or use only distant and indirect indicators. Indicators of access and completion are quite specific. Yet, on most access indicators, even the low-income countries are already achieving at levels that are increasingly close to the levels in the developed world; enrollment rates in primary school, even in poor countries, are at around 80% to 90% of what they are in rich countries. This holds true for Kenya. The gross enrollment ratio at the primary level in Kenya is already at 107.2% and the net ratio is a respectable 83.2%.[4] Learning levels are a different matter. Children (enrolled children, at that) in poor countries, including Kenya, learn about 30% as much as children in rich countries or, to put it another way, the average child in a poor country learns at about the same level as the poorest-performing 3% of children in rich countries. (The exact data for Kenya would probably put this average child at levels somewhat above the poorer countries.)

[4] But recall that the net enrollment rate is a "noisy" measure of access, as it includes measures of internal efficiency.

Noting this situation, government officials, scholars, and to a lesser degree activists, are putting increasing pressure on government programs and donor and technical assistance agencies to track outcomes, and to prove that education spending is leading to more learning by children. Many existing programs such as TIMSS, PISA, PIRLS, and SACMEQ[5] provide an invaluable service in allowing countries to compare themselves to each other. In addition, various agencies have been working on assessments, such as the EGRA, that allow a very early diagnosis of reading issues. This is done under two or three hypotheses:

- First, that the learning problems detected in various written assessments such as SACMEQ or PISA are rooted in problems that start in grade 1 or even before. Yet, SACMEQ and PISA are administered in later grades or age groups.
- Second, that assessing children's reading ability using oral measurements can provide clues as to how to improve early performance.
- Third, that simple measures that focus on orality and very basic skills are easy to stream into teachers' practices, because their implications are relatively obvious.

In 2006, RTI and AKF opened up discussions to first assess the level of reading of children in Malindi district as a baseline, and then, depending on what was found, progress to develop an intervention. These discussions involved, as well, quality assurance, examinations, and information systems personnel from the Ministry of Education. The baseline assessment was carried out in June 2007. The findings showed low levels of letter-naming ability (measured in terms of correct letter-naming fluency, per minute) and ability to fluently read and comprehend simple passages. The assessments found more or less equal lack of skills in Kiswahili and English. (The same children were tested in both languages—also an important innovation—and the correlation between children's skills across languages was found to be very high.)

In response, an intervention was designed, and was applied to 20 "treatment" schools from within the sphere of the EMACK II project. In addition, 20 "control" (mostly non-EMACK) schools were also selected.[6] The selection of schools into those groups was random, although treatment schools were all already involved in the EMACK II project. The treatment was fairly simple, and consisted, essentially, of a set of very tightly designed lessons in reading that focused on the skills that research shows are essential to a speedy uptake of reading: phonological awareness (pre-reading skills, including listening and sound sensitivity), alphabetic principle (relationship of print to sound), vocabulary, fluency, and comprehension. Specific, lesson-by-lesson, week-

[5] Trends in International Mathematics and Science Study (TIMSS); Organisation for Economic Co-Operation and Development's Programme for International Student Assessment (PISA); Progress in International Reading Literacy (PIRLS); Southern Africa Consortium for the Measurement of Educational Quality (SACMEQ).

[6] In that sense there is a bit of an identification problem, in that the "treatment" really consisted, on the whole, of the EMACK II interventions, not just the reading intervention. However, the fact that at baseline the EMACK II schools were reading no better than the control schools (a little worse, in fact) suggests that this combination of comparison group and baseline-plus-post-treatment measures can isolate the treatment effect. It would have been ideal to randomize over the entire population, but this was not practical.

by-week, lesson plans were developed, and were disseminated to the schools. These were made to fit within a clear and explicit scope and sequence for the whole year.

The teachers in grades 1 and 2 were trained over five-day period on how to use these methods. They also agreed to teach three reading lessons per subject each week in place of language instruction. The lessons were developed for both Kiswahili and English, and at the development stage it was ensured that they all were in line with the Kenyan curriculum. The lessons were extended, in theory, only to the treatment schools. In addition, the treatment schools received visits from supervisors, and support from teacher trainers. Both the control and treatment schools were assessed formally, as noted, both at the beginning and at the end of the process. In addition, a varying subset of the schools was assessed informally during the year.

Only one grade, grade 2, was targeted, although the grade 1 teachers were trained and also implemented the program in the treatment schools. As noted, the EGRA Kenya project was implemented in the Malindi district. With over a half a million inhabitants, Malindi is one of the 13 administrative districts of the Coast Province of Kenya, and is considered to be among the poorest districts in Kenya. The district is further divided into seven zones, through which a total of 120 primary schools are served.

Out of the district's 120 primary schools, 40 were selected to become part of the EGRA Kenya study. These 40 schools were selected at random by RTI's implementing partner, AKF. Out of these 40 schools, 25 were part of AKF's schools targeted through EMACK II, and 15 were not. There was some concern to create some physical distance between treatment and control schools, yet since the experiment was within the Malindi District itself, in the end the control schools were selected so as to be relatively separate from the treatment schools but within the same district. This meant that some of the control schools were along the coastal strip, dominated by tourist hotels, with many of the parents being better off (in terms of their socioeconomic status and education) than those in the rural part of the district. (This whole point is the subject of a discussion below dealing with the issue of contamination of practices, and the pitfalls created by contamination but also by mechanisms that avoid contamination via physical separation.)

The control schools would receive no intervention; the treatment schools would receive teacher training and support in implementing teaching practices to improve the teaching of reading. As will be seen below, the "density" of the 40 (20 treatment and 20 control) schools over the 120 in the district has important implications: a third of the schools in the district were being assessed.

The division of labor among RTI, AKF, and a third-party survey firm, East Africa Development Consultants (EADEC), was also an interesting feature of the program. The division of labor was as follows:

- RTI appointed EADEC to conduct the baseline assessment in the selected control and treatment schools. RTI, however, led the process of adjusting the EGRA instruments to Kenyan context during a stakeholder workshop.
- RTI designed the remedial intervention framework—that is, a scope and sequence and lesson plans for teaching reading in English and Kiswahili, as

well as draft daily lesson plans. This process was facilitated by RTI's expert, with instrumental help from local Kenyan experts, including those of AKF.

- AKF finalized the remedial intervention design, trained teachers in treatment schools in grades 1 and 2, and provided ongoing support to these teachers during one academic year.
- AKF conducted two informal assessments to determine if students were making any progress on reading performance.
- RTI and EADEC conducted the post-treatment assessment at the end of the project.
- The district education office extended its support to this project as well. The AKF and district education officers jointly supported schools as well as conducting informal assessments.

The original schedule for the project's implementation was to conduct a baseline assessment in June 2007, use the baseline findings to inform the remedial intervention design, and commence the intervention in the September or so of 2007. There was a deviation from this schedule in the sense that the intervention did not start until February 2008 due AKF's need to secure additional funding.

Furthermore, to understand the materials below, it is important to note that the basic school calendar consists of three terms interspersed with one-month vacations between terms. The terms are January–March, May–July, and September–November, with breaks, essentially, in April, August, and December. It is important also to note that the 2008 school year was topsy-turvy in Kenya due to the political violence. The baseline measurement was carried out at the end of July 2007. The post-treatment assessment was carried out at the end of November 2008. Given all these factors, it seems valid to say that the children who were tested in November 2008 had had somewhere between one and two more months of actual instruction than those who were tested in July of 2007. We will assume, to be conservative, that the children had had two more months of instruction. These facts need to be recalled in what follows.

In all that follows it is also important to keep in mind the fact that the project was meant not only to create some learning around reading per se, but also to have a capacity-building effect on Kenyan colleagues. In various cases, for example, procedures may not have been implemented with the intensity or accuracy that is ideal. But it is difficult for partners to realize the importance of some of these things before getting hard measurement results. To some degree, the purpose of capacity-building on measurement, and the use of measurement in tracking, would help create awareness of the importance of accurate measurement, and also the importance of intensive implementation. But for this importance to come through, one has to see the results. We are now at that stage. The process, it is safe to conclude, has led to a great deal of learning on the part of all partners, and has upped the understanding of all concerned regarding the value of rigorous interventions (in reading), and rigorous measurement, including qualitative or "forensic" follow-ups.

3. The Quantitative Results

3.1 Instruments and calibration of instruments

(This section contains some technical detail. The reader may skip it, or come back to it after focusing on the more substantive results that follow.)

The instruments used for the post-treatment assessment are identical to those of the June 2007 assessment in terms of their format and components, but not 100% equivalent in terms of content of each individual component, particularly in the area of connected text. As noted, all of the subtests or components used in the baseline were also administered in the post-treatment assessment. But their exact content needed to be changed, to prevent "teaching to the test" situations. RTI did not have control over how widely, and if at all, the baseline assessment tools were shared and we did not want to risk a possibility of children memorizing words, passages, and answers to questions. Yet, in assessing possible improvement in an early grade reading project, it is important to calibrate the pre- and post-tests to make sure they are of equal difficulty or, if not of exactly equal difficulty, to be able to "translate" results in one to make them equivalent to results in the other.

Instruments for both baseline and post-treatment assessment can be provided upon request to the authors. The only change—or, rather, addition—from the baseline to the treatment instruments (other than changes needed to provide equivalent-difficulty but different items) was a set of questions that will inform future studies that could look into the links between reading performance and various health factors. Answers to these questions were not analyzed in this report. Depending on requests and need, they may be analyzed in subsequent versions of this report.

As for other components of the instrument (both English and Kiswahili instruments), the following was performed: Letters were randomly reshuffled so that they appeared in a different order in the post-test than in the pre-test (to prevent improved performance via memorization), words were also randomly reshuffled for the same reason, some new words were used to replace the old words (they are all still high-frequency words appropriate for grade 2), and new passages were developed for both languages to test fluency and comprehension in reading connected text. Finally, for the phonemic awareness task in the English language, some of the words that were used in the baseline were kept, while others were added. They were ordered, however, from easy to hard, so that the difficulty of this task would be identical to the one in the baseline.

In the connected text passages and their adjustments, we took care to ensure that the newly developed passages were equal in length and difficulty to the old passages used in the pre-test, or were mathematically equated. This is an important aspect from the measurement point of view and needs to be explained carefully. The first step in equating old and new passages was to ensure that (a) they were of approximately equal length, and (b) the words used were frequently used words. This needs to be explained a bit further.

Reading connected text is a timed exercise and for developing countries in which EGRA is being implemented, and in the grades being assessed, EGRA experts have agreed that a benchmark speed at which students should read is about one word per second. From this point of view, it is important that the new passage take about the same time as the old passage, so that the students do get the same task at the time of testing for both baseline and assessment. The number of words is fairly easy to control and keep constant.

Level of reading difficulty is a slightly more complex matter. In this matter the Kenyan colleagues were perhaps not as easily convinced of the importance of maintaining constancy. Rather than waste time on debates, and given that the importance of certain issues is easier to demonstrate with data, we agreed that approximate constancy was sufficient, and that we could equate using ex-post analyses, as follows.

First, prior to the training of assessors and deployment, the EGRA team administered the newly developed passages for both languages to about 20 students that were independent of the post-test sample. There were two assessors, each given 10 students. One assessor was tasked to ask students to read the old passage first, and then the new passage. The other assessor reversed the order. Students were timed and it was found the new passages demanded more time to be read than the old passages. On this basis some initial adjustments were made. We consulted a local district quality assurance officer to assist with some problematic words in the new passage and offer alternatives. For instance, in the first draft of the post-test English passage the word “seashore” was used. On the basis of local advice, it was agreed to use “beach,” given that children would more likely use the word “beach” than “seashore.”

Yet, in order to be completely sure and to have sufficient data for ex-post calibration or equating, another test was conducted. An independent sample of children was again tested following the same process described above and the averages confirm that the two passages of the post-treatment instrument were slightly more difficult in both English and Kiswahili. Table 1 shows the results.

Table 1. Time (seconds) in reading connected text passage

English	Pre	70.5
	Post	75.25
	Difference	4.75
Kiswahili	Pre	93.85
	Post	97.3
	Difference	3.45

This means that it was necessary to carry out some ex-post calibration on the reading fluency data. In doing this calibration, several issues arise. First, is there a multiplicative and not merely additive difference between the two? Second, what is the size of the difference, if multiplicative? And, third, is the difference “reliable” so that calibration is valid? To investigate this, we carried out a simple regression

between the post-test as a right-hand variable and the pre-test as a left-hand variable. This yielded the results in Box 1 in English and Kiswahili.

Box 1. Pre- and post-treatment passage calibration

```
 Source | SS df MS Number of obs = 20
-------------+------------------------------ F( 1, 18) = 75.56
 Model | 10552.9885 1 10552.9885 Prob > F = 0.0000
 Residual | 2514.01153 18 139.667307 R-squared = 0.8076
-------------+------------------------------ Adj R-squared = 0.7969
 Total | 13067 19 687.736842 Root MSE = 11.818

------------------------------------------------------------------------------
 engold | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
 engnew | .9547192 .1098336 8.69 0.000 .7239673 1.185471
 _cons | -1.342618 8.677169 -0.15 0.879 -19.57267 16.88744
------------------------------------------------------------------------------



 Source | SS df MS Number of obs = 20
-------------+------------------------------ F( 1, 18) = 246.51
 Model | 18825.9021 1 18825.9021 Prob > F = 0.0000
 Residual | 1374.64788 18 76.3693267 R-squared = 0.9319
-------------+------------------------------ Adj R-squared = 0.9282
 Total | 20200.55 19 1063.18684 Root MSE = 8.739

------------------------------------------------------------------------------
 kiswold | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
 kisnew | 1.061039 .0675792 15.70 0.000 .9190601 1.203017
 _cons | -9.389058 6.859667 -1.37 0.188 -23.80068 5.022569
------------------------------------------------------------------------------
```

Given that neither of the constant terms is statistically significant, each of the regressions was re-run eliminating the constant term, to yield the results shown in Box 2.

Box 2. Pre- and post-treatment passage calibration, zero intercept

```
 Source | SS df MS Number of obs = 20
-------------+------------------------------ F( 1, 19) = 829.89
 Model | 109954.645 1 109954.645 Prob > F = 0.0000
 Residual | 2517.35536 19 132.492387 R-squared = 0.9776
-------------+------------------------------ Adj R-squared = 0.9764
 Total | 112472 20 5623.6 Root MSE = 11.511

------------------------------------------------------------------------------
 engold | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
 engnew | .9385319 .032579 28.81 0.000 .8703433 1.006721
------------------------------------------------------------------------------



 Source | SS df MS Number of obs = 20
-------------+------------------------------ F( 1, 19) = 2439.15
 Model | 194839.279 1 194839.279 Prob > F = 0.0000
 Residual | 1517.72061 19 79.8800322 R-squared = 0.9923
-------------+------------------------------ Adj R-squared = 0.9919
 Total | 196357 20 9817.85 Root MSE = 8.9376

------------------------------------------------------------------------------
 kiswold | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
 kisnew | .9723732 .0196886 49.39 0.000 .9311646 1.013582
------------------------------------------------------------------------------
```

The results are convincing that there is a strong multiplicative relationship, not merely an additive relationship (in fact the relationship is not really additive at all). And the relationship is very strong. This means that prior to analyzing the results, which is what was done for this report, the post-tests had to be adjusted as follows, in terms of the time they took:

- Adjusted English post-test time = 0.938 * English post-test time
- Adjusted Kiswahili post test time = 0.972 * Kiswahili post-test time

Or, they can be adjusted as follows in terms of correct words per minute:

- Adjusted English post-test correct words per minute = 1.066 * English correct words per minute
- Adjusted Kiswahili post-test correct words per minute = 1.0288 * Kiswahili correct words per minute

This is sufficient to carry out a recalibration of the post-test. However, to confirm the analysis, we also carried out a basic readability analysis, using two versions of a Spache analysis and one version of the Dale-Chall analysis to determine the levels of difficulty for both passages. This analysis was performed only for the English-language passage since no such tools exist for the Kiswahili language, to our knowledge. The following are the old and new passages and difficulty levels for English only. As can be seen, the new English passage appears to be some 20% difficult than the old passage—although these reading difficulty assessments are hardly exact. Thus, our mathematical calibration, shown above, is, if anything, a bit conservative. The post-test passage was clearly slightly more difficult. In subsequent work of this sort in Kenya it is important that researchers and persons involved in monitoring and evaluation realize the value of very accurate calibration and equating of the passage difficulty. In any case, the post-test was certainly no easier than the pre-test. This lends credence to the results.

English language

Old passage

Kazungu had a little dog. The little dog was fat. One day Kazungu and the dog went out to play. The little dog got lost. But after a while the dog came back. Kazungu took the dog home. When they got home Kazungu gave the dog a big bone. The little dog was happy so he slept. Kazungu also went to sleep.

[Note: to prevent the readability analysis from artificially increasing the difficulty of the passage due to the unfamiliar name "Kazungu," this name was replaced by an English name of equivalent length, "Jonathan." Thus, the passage analyzed was: Jonathan had a little dog. The little dog was fat. One day Jonathan and the dog went out to play. The little dog got lost. But after a while the dog came back. Jonathan took the dog home. When they got home Jonathan gave the dog a big bone. The little dog was happy so he slept. Jonathan also went to sleep.]

Spache analysis from Okapi:[7]
Total words in sample: 62
Total sentences in sample: 9
Average number of words per sentence: 6.88
Number of words not matched to revised Spache word list: 2
Percentage of words not matched to revised Spache word list: 3.22

[7] http://www.interventioncentral.org/htmdocs/tools/okapi/okapi.php

Spache readability index: 2.08

Spache readability index from Micro Power and Light: 2.2

Dale-Chall readability index from Okapi: 3.97

New passage:
Tom was on holiday at the beach. He loved to play on the sands. One day he was tired and sat on a chair. He saw a ship coming into the bay. Tom knew it was an enemy ship. Soon a small boat came to the shore. Three thieves jumped out, caught Tom, and took him back to the ship.

Spache from Okapi:
Total words in sample: 61
Total sentences in sample: 7
Average number of words per sentence: 8.71
Number of words not matched to revised Spache word list: 3
Percentage of words not matched to revised Spache word list: 4.91
Spache readability index: 2.49

Spache readability index from Micro Power and Light: 3.2

Dale-Chall readability index from Okapi: 4.32

Kiswahili language

Old passage
Jumamosi iliyopita Katana na dada zake, Kadzo na Fatuma, walienda kuogelea baharini. Kabla ya kuondoka walibeba mahamri, maembe, samaki na maji ya machungwa.
Walibeba pia nguo zao za kuogelea. Wote waliingia kwenye matatu kuelekea huko. Walipofika baharini waliona watu wengi sana. Katana alikuwa na hamu sana ya kuogelea. Maskini Katana, aliingia baharini bila kubadili nguo zake! Dada zake walimcheka sana.

New passage:
Usiku wa manane Kalume na dadake walisikia sauti ya motokaa inakaribia na wakatoka nje. Wakasimamisha motokaa hiyo. Mwenye gari aliwapatia lifti. Baada ya kusafiri kwa muda mrefu Mzungu aliwateremsha karibu na kibanda. Hapo nje ya kibanda palikuwa na moto. Kulipokuwa karibu kucha, Sidi alisema kwamba alikuwa akiumwa na wadudu. Hata kalume pia alisema vivyo hivyo. Wakatoka nje ya kibanda mbio mbio.

3.2 Results

We present a summary of the results before delving into the details.

1. **Student background questions.** The data on background questions confirmed that the socioeconomic status of students and conditions in which they were living and studying were not different in the post-test from the baseline. Essentially the same populations were being tested.
2. **Overall performance in treatment and control schools.** Without differentiating between control and treatment schools, significant improvements took place on almost all tasks in both languages. From this vantage point, it can be said that the experiment worked very well.
3. **Performance by school type – control vs. treatment schools**. The results show that the control schools performed about as well as the treatment schools. Given these findings and their huge implications for this project, but

also future projects of this kind, it is important to further explore the reasons as to why the control schools improved as much as the treatment schools (in absolute terms—in percentage terms the treatment schools improved more). This report shows the results of these explorations.

4. **Did the intervention work?** The conclusion is that the treatment yielded intended results. But the conclusions regarding leakage of the treatment to the control schools is even more interesting, and it is discussed in detailed below.

3.2.1 Student background questions

When compared, baseline and post-treatment data showed that the population of students in the pre- and post-test was the same. This finding ruled out a possibility that the students in the post-test were doing better in reading because, by luck of the draw, they might have come from a different social class. In the case of language spoken in the home, the data are almost identical for both baseline and post-treatment tests, as well as when compared by school type—Table 2 for the baseline data and Table 3 for the post-treatment data. In Table 4 we have data on various questions: pre-primary education attendance, availability of reading materials at home, and availability of TV or radio. With the exception of the availability of reading material at home, there were no significant differences between baseline and post-treatment assessment.

A difference can be noted on the availability of reading materials at home, where students at the time of the baseline seemed to have had more reading material available at home than at the time of the post-treatment assessment. One cannot tell whether this was a real change in circumstances, or was due to some slight change in how the questions might have been asked, or reflected a slightly different population. In any case, correlations between these two variables and student reading performance are not substantive, which leads us to conclude that the difference in any case would not have had any impact on the measured changes in reading. The correlation in Box 3 shows this point.

Box 3. Correlation reading materials and reading fluency

```
Baseline
. correl r_mat w_cor
(obs=799)

 | r_mat w_cor
-------------+------------------
 r_mat | 1.0000
 w_cor | -0.1341 1.0000

Post-treatment
. correl read_mat paspermintimediffadj
(obs=800)

 | read_mat pasper..
-------------+------------------
 read_mat | 1.0000
paspermint~j | -0.1262 1.0000
```

Table 2. Language spoken at home: Percentage of students by language spoken and home, control and treatment schools, baseline

	Treatment		Control		Total	
	Freq	Perc	Freq	Perc	Freq	Perc
English	2	0.5	2	0.5	4	0.5
Kiswahili	61	15.3	73	18.3	134	16.8
Other	337	84.3	325	81.3	662	82.8

Calculated by the authors

Table 3. Language spoken at home: Percentage of students by language spoken at home, control and treatment schools, post-treatment

	Treatment		Control		Total	
	Freq	Perc	Freq	Perc	Freq	Perc
English	1	0	2	1	4	0
Kiswahili	36	9	52	13	134	11
Other	363	91	346	87	662	89

Calculated by the authors

Table 4. Percentage of students who responded "yes" on background questions; overall and compared by school type

	Baseline		Post-treatment	
	treatment	control	treatment	control
Preschool education attendance	91.5	93	96	93
Assistance with homework	86	88	90	85
Reading material available at home	75	81	54	59
Do you watch TV	10	20.5	10	21
Do you listen to radio	69.5	70	68	71

Calculated by the authors

3.2.2 Overall performance – both treatment and control schools

Looking at average performance on all tasks, it can be noted that significant improvements took place between baseline and post-treatment assessment. Exceptions are comprehension scores and phonemic awareness in English. Table 4 shows comparisons in average performance on baseline and post-treatment tests as well as changes in absolute and percent terms. Given that the EGRA Kenya project was an experiment through which we aimed to determine whether the intervention had worked or not, we must look at the scores by control and treatment schools. The results are rather interesting from the experimental point of view—see next section and Table 5.

Table 5. Performance on all tasks by all schools (treatment and control)

Language and task	Baseline Average	Post-treatment Average	Absolute Change	Percent change
Kiswahili				
Letter recognition	4.7	20.6	15.9	338%
Word recognition	11.7	20.8	9.1	78%
Passage words	10.2	18.9	8.7	85%
Comprehension score	0.4	0.5	0.1	25%
English				
Letter recognition	22.7	29.5	6.8	30%
Word recognition	7.5	16	8.5	113%
Passage words	11.4	20.85	9.45	83%
Comprehension score	0.4	0.3	-0.1	-25%
Phoneme segmentation	11.5	10.9	-0.6	-5%

Calculated by the authors

The (negative) difference between baseline and post-test in comprehension is not significant, and may have to do with very slight differences in how the test was marked. More detailed analysis could be done on this, but in any case the differences are not significant. The most important aspect is the very large set of differences in other skills. Below we note some reasons for concern over project impact, in spite of these big differences. But it is interesting that one of the areas the project tried to impact was Kiswahili letter naming recognition and fluency. In the baseline this was very poor: much worse in Kiswahili than in English. This was noted and it was suggested that this skill should be worked on. The impact is notable. Whereas in the post-test, letter recognition and letter-naming fluency were still not as good in Kiswahili as in English, they did improve hugely. The fact that one would see more improvement in a weak area, and one that was emphasized in the project, suggests project impact. But, in general, if viewed based on this table, the project seems to have had a strong impact. However, a more detailed analysis reveals some ambiguities and puzzling patterns.

3.2.3 Performance by school type—Control vs. treatment schools

The main issue to be confronted is that control schools appear to have improved about as much as the treatment schools. This can be seen in all areas of the assessment, as shown in Table 6 and Table 7 (Table 7 shows the same information as Table 6, but shows, in addition, the percentage changes from baseline to post-test). The impact, as already shown, appears to be large, but appears basically the same in both types of schools. However, note that the effect sizes and the statistical significance of the baseline-to-post-treatment difference tends to favor the treatment schools, particularly in some of the more important skills, such as passage reading fluency. The standard deviation for the effect size is pooled. We did not calculate these factors for the

comprehension scores, as they clearly did not improve and this can be, perhaps, attributed to differences in measurement pre- and post-treatment).

Table 6. Performance on all tasks by control and treatment schools

		Kiswahili				English			
		Baseline	Post-treatment	Effect size	*p* value of diff	Baseline	Post-treatment	Effect size	*p* value of diff
Letter recognition	T	4.8	20.9	.42	.0000	21.6	29.6	.21	.0016
	C	4.5	20.3	.51	.0000	23.8	29.4	.13	.0382
Word recognition	T	10	19.6	.37	.0000	5.8	13.6	.34	.0000
	C	13.3	22	.27	.0001	9.1	18.4	.25	.0002
Passage reading	T	8.7	17.4	.35	.0000	9.3	18.3	.27	.0001
	C	11.8	20.4	.27	.0001	13.4	23.4	.21	.0018
Comprehension questions	T	0.36	0.74			0.34	0.27		
	C	0.53	0.32			0.45	0.37		

C = control; T = treatment

Table 7. Performance on all tasks by control and treatment schools (includes absolute and percent changes)

		Baseline	Post-treatment			Baseline	Post-treatment		
		Kiswahili	Kiswahili			English	English		
		Average	Average	Absolute change	Percent change	Average	Average	Absolute change	Percent change
Letter recognition	T	4.8	20.9	16.1	335%	21.6	29.6	8	37%
	C	4.5	20.3	15.8	351%	23.8	29.4	5.6	24%
Word recognition	T	10	19.6	9.6	96%	5.8	13.6	7.8	134%
	C	13.3	20	6.7	50%	9.1	18.4	9.3	102%
Passage reading	T	8.7	17.4	8.7	100%	9.3	18.3	9	97%
	C	11.8	20.4	8.6	73%	13.4	23.4	10	75%
Comprehension questions	T	0.36	0.74	0.38	106%	0.34	0.27	-0.07	-21%
	C	0.53	0.32	-0.21	-40%	0.45	0.37	-0.08	-18%

C = control; T = treatment

3.2.4 A more nuanced take on improvements

A school-by-school analysis, shown in Table 8 and Table 9, as opposed to simply comparing averages in the two groups, suggests that the impact of the treatment is even more noticeable in certain schools, something that is somewhat hidden in the averages. This is shown in some of the highlights. Importantly, it seems logical that if some schools improved so much more than others, then it is unlikely that the improvement in both control and treatment schools was due to a generalized upward drift in all schools due to some unobserved third factor.

A powerful third factor, such as perhaps better distribution of textbooks in the whole district, would not seem likely to affect some schools so much than others. A school-by-school analysis also makes it possible to focus the analysis a bit on the outlier

schools, particularly those that were doing really badly before the intervention, to see what happened to them: Did the effort bring up the worst-performing at the baseline? The graphics following the tables (see Figure 1) are a bit challenging but make the point quite well. The graphics are used for the English portion of the assessment only, and only for two variables: familiar word fluency and connected text fluency. The graphics show, on the horizontal axis, the baseline scores of the schools, and on the vertical axis the post-test scores. The scatter of points represents all the schools. The diagonal line shows the "line of no difference." If schools had simply not changed at all, all of them would be on the diagonal line. The fact that almost all schools are above the diagonal lines for both skills (words and connected text) means that almost all schools improved on both skills.

Table 8. Individual listing of schools—Kiswahili[8]

School Name	Baseline School Type	Baseline Kswh. words	Post-treatment Kswh. Words	Absolute change Kswh. words	Percent change Kswh. Words	Baseline Kswh. Passage words	Post-treatment Kswh. Passage words	Absolute change Kswh. Passage words	Percent change Kswh. Passage words
School A	Ctrl	22.5	28.7	6.2	28%	24.3	26.9	2.6	11%
School B	Ctrl	20.1	31.5	11.4	56%	18.0	31.4	13.4	75%
School C	Ctrl	7.6	14.4	6.8	89%	7.3	15.9	8.7	120%
School D	Ctrl	1.2	6.3	5.1	425%	0.9	4.9	4.0	475%
School E	Ctrl	14.2	24.0	9.8	69%	12.0	18.9	6.9	58%
School F	Ctrl	8.1	22.6	14.5	179%	5.8	18.4	12.6	218%
School G	Ctrl	9.4	17.4	8.0	86%	6.8	16.6	9.8	144%
School H	Ctrl	10.8	16.5	5.7	52%	10.7	16.6	5.9	56%
School I	Ctrl	12.4	17.9	5.5	45%	11.2	17.1	5.9	53%
School J	Ctrl	6.4	8.6	2.3	35%	4.4	7.6	3.2	74%
School K	Ctrl	7.2	13.5	6.3	88%	6.0	13.3	7.3	121%
School L	Ctrl	13.1	20.2	7.2	55%	11.7	17.8	6.2	53%
School M	Ctrl	24.9	36.8	11.9	48%	24.6	32.7	8.1	33%
School N	Ctrl	27.0	27.6	0.6	2%	26.1	28.4	2.3	9%
School O	Ctrl	11.5	25.0	13.5	117%	9.8	21.4	11.6	118%
School P	Ctrl	13.1	20.7	7.6	58%	11.9	18.5	6.7	56%
School Q	Ctrl	15.6	25.3	9.7	63%	13.5	23.9	10.4	77%
School R	Ctrl	16.6	22.9	6.3	38%	12.0	23.1	11.1	93%
School S	Ctrl	9.4	27.3	17.9	190%	7.1	25.9	18.9	268%
School T	Ctrl	15.5	33.8	18.3	118%	12.4	29.0	16.6	135%
School U	Trtmnt	6.4	7.0	0.6	9%	4.7	7.0	2.3	50%
School V	Trtmnt	7.5	17.9	10.5	140%	6.6	18.2	11.6	175%
School W	Trtmnt	8.9	30.0	21.1	237%	6.7	26.7	20.1	302%
School X	Trtmnt	5.3	21.5	16.2	309%	4.4	20.5	16.1	371%
School Y	Trtmnt	8.0	16.4	8.4	104%	6.6	15.1	8.5	129%
School Z	Trtmnt	10.3	12.3	2.0	20%	8.0	11.0	3.0	38%
School AA	Trtmnt	8.2	26.1	18.0	221%	6.9	23.6	16.8	245%
School AB	Trtmnt	21.5	24.6	3.2	15%	20.3	20.3	0.0	0%
School AC	Trtmnt	10.9	15.8	4.9	45%	7.9	11.3	3.5	44%
School AD	Trtmnt	17.9	23.1	5.3	30%	23.1	20.0	-3.0	-13%
School AE	Trtmnt	15.9	27.5	11.6	73%	11.7	23.4	11.8	101%

[8] Schools have been made anonymous to protect their privacy.

School Name	Baseline School Type	Baseline Kswh. words	Post-treatment Kswh. Words	Absolute change Kswh. words	Percent change Kswh. Words	Baseline Kswh. Passage words	Post-treatment Kswh. Passage words	Absolute change Kswh. Passage words	Percent change Kswh. Passage words
School AF	Trtmnt	0.2	14.0	13.8	6898%	0.0	12.2	12.2	NA
School AG	Trtmnt	4.1	20.0	15.9	393%	2.8	17.1	14.4	523%
School AH	Trtmnt	9.1	18.0	8.9	98%	8.8	14.0	5.3	60%
School AI	Trtmnt	14.5	14.0	-0.6	-4%	11.9	11.3	-0.6	-5%
School AJ	Trtmnt	17.3	29.7	12.4	72%	15.9	23.5	7.6	48%
School AK	Trtmnt	3.8	22.9	19.1	504%	2.3	22.0	19.7	855%
School AL	Trtmnt	11.8	22.6	10.8	92%	9.3	22.8	13.5	145%
School AM	Trtmnt	9.8	11.7	1.9	19%	7.2	15.1	8.0	112%
School AN	Trtmnt	9.1	17.0	7.9	87%	9.5	13.6	4.1	43%

Table 9. Individual listing of schools—English[9]

School Name	Baseline School Type	Baseline Eng. Words	Post-Trtmnt Eng. Words	Absolute change Eng. Words	Percent change Eng. Words	Baseline Eng. Passage words	Post-Trtmnt Eng. Passage words	Absolute change Eng. Passage words	Percent change Eng. Passage words
School A	Ctrl	24.6	27.7	3.1	12%	36.2	31.3	-4.9	-13%
School B	Ctrl	12.0	27.9	16.0	134%	18.5	40.1	21.6	117%
School C	Ctrl	6.0	11.9	5.9	98%	7.5	14.8	7.3	98%
School D	Ctrl	0.6	5.9	5.3	875%	1.7	5.3	3.7	223%
School E	Ctrl	7.3	18.6	11.3	155%	10.0	22.6	12.6	126%
School F	Ctrl	5.0	17.5	12.6	254%	6.4	20.5	14.1	222%
School G	Ctrl	4.4	10.2	5.8	131%	6.8	14.9	8.2	121%
School H	Ctrl	5.6	10.9	5.4	96%	8.6	9.6	1.0	12%
School I	Ctrl	6.8	9.9	3.1	46%	10.8	16.5	5.6	52%
School J	Ctrl	3.9	4.0	0.2	4%	6.1	4.7	-1.3	-22%
School K	Ctrl	3.2	10.2	7.0	217%	3.8	14.7	11.0	292%
School L	Ctrl	8.1	14.6	6.5	80%	10.5	19.4	8.9	85%
School M	Ctrl	24.7	39.1	14.5	59%	30.3	47.4	17.2	57%
School N	Ctrl	21.4	31.0	9.6	45%	31.2	42.2	11.1	36%
School O	Ctrl	4.7	14.7	10.0	213%	8.6	17.4	8.8	102%
School P	Ctrl	8.0	15.8	7.8	97%	14.1	22.8	8.8	62%
School Q	Ctrl	9.5	33.1	23.6	248%	14.9	35.7	20.8	140%
School R	Ctrl	9.8	18.5	8.7	89%	16.4	25.0	8.6	53%
School S	Ctrl	6.2	19.1	12.9	208%	11.1	28.5	17.4	157%
School T	Ctrl	11.0	27.7	16.8	153%	15.7	34.5	18.8	120%
School U	Trtmnt	3.9	3.2	-0.7	-17%	6.2	4.6	-1.6	-25%
School V	Trtmnt	4.3	14.2	9.9	233%	7.5	23.6	16.1	215%
School W	Trtmnt	3.4	25.0	21.6	634%	8.5	35.2	26.8	317%
School X	Trtmnt	2.2	12.5	10.3	466%	3.1	14.2	11.1	365%
School Y	Trtmnt	2.6	8.9	6.3	240%	6.3	13.5	7.3	117%
School Z	Trtmnt	4.6	10.3	5.7	125%	6.8	10.9	4.1	61%
School AA	Trtmnt	3.1	19.2	16.1	518%	5.5	24.1	18.7	343%
School AB	Trtmnt	12.7	12.4	-0.3	-2%	21.2	14.2	-7.0	-33%
School	Trtmnt	5.4	8.7	3.3	62%	9.5	11.6	2.1	22%

[9] Schools have been made anonymous to protect their privacy.

School Name	Baseline School Type	Baseline Eng. Words	Post-Trtmnt Eng. Words	Absolute change Eng. Words	Percent change Eng. Words	Baseline Eng. Passage words	Post-Trtmnt Eng. Passage words	Absolute change Eng. Passage words	Percent change Eng. Passage words
AC									
School AD	Trtmnt	14.4	19.1	4.7	33%	23.9	26.2	2.3	10%
School AE	Trtmnt	11.7	19.8	8.1	69%	15.1	27.0	11.9	79%
School AF	Trtmnt	0.2	14.0	13.8	6903%	0.3	14.5	14.3	5708%
School AG	Trtmnt	0.7	9.6	8.9	1264%	2.0	15.7	13.8	706%
School AH	Trtmnt	2.2	12.7	10.5	475%	6.1	20.0	13.9	230%
School AI	Trtmnt	8.9	6.8	-2.1	-23%	10.6	7.2	-3.3	-31%
School AJ	Trtmnt	13.4	22.4	9.1	68%	21.6	30.0	8.4	39%
School AK	Trtmnt	1.6	19.1	17.6	1133%	2.1	22.9	20.8	991%
School AL	Trtmnt	8.9	18.6	9.8	110%	12.4	27.1	14.8	120%
School AM	Trtmnt	7.4	9.3	2.0	27%	9.6	13.4	3.8	40%
School AN	Trtmnt	5.0	7.6	2.6	53%	8.3	10.0	1.7	21%

The graphics are not shown for the same two skills in Kiswahili, but they would look the same. Note that each school is labeled with a C for "control" or a T for "treatment." The fact that for both skills the schools toward the upper right of the scatter of points are control schools means that there were more very good (relatively speaking) control schools at the baseline, and these got better. The fact that the lower left hand of the scatter of points tends to contain more treatment schools means that the very "worst" schools at the baseline happened to be treatment schools. These also improved, quite a lot, so the effort does seem to have brought up the schools that were worst off in the baseline.

Now, the graphics are presented both on a logarithmic scale and a linear scale (logarithmic first, linear second, in each case). Using a logarithmic scale means that the distance from the dot representing any school to the "line of no difference" represents the *percentage* difference between the baseline and the post-treatment. For the linear scale, the distance from the dot to the diagonal line of no difference represents the absolute change. It is clear from both cases that the treatment schools, which tended to be slightly lower-achieving in the baseline (or, perhaps more accurately, tended to have a slightly higher concentration of very poorly performing schools in the baseline), tended to then improve more, certainly in percentage terms. The change in the most poorly performing schools at baseline is particularly noteworthy, and it turns out that many of these were in the treatment group. Thus, this finding adds to the sense of project impact, in this important respect.[10]

[10] Unfortunately there are not enough schools in the sample to do a very reliable statistical analysis at the school level. (But it can be done at the student level, yet at the student level one cannot observe the growth, due to the nature of the experiment.) Nonetheless, it is possible to assert that the percentage increase was statistically significantly better for the treatment schools, but only at the 10% level, for Kiswahili connected text and English

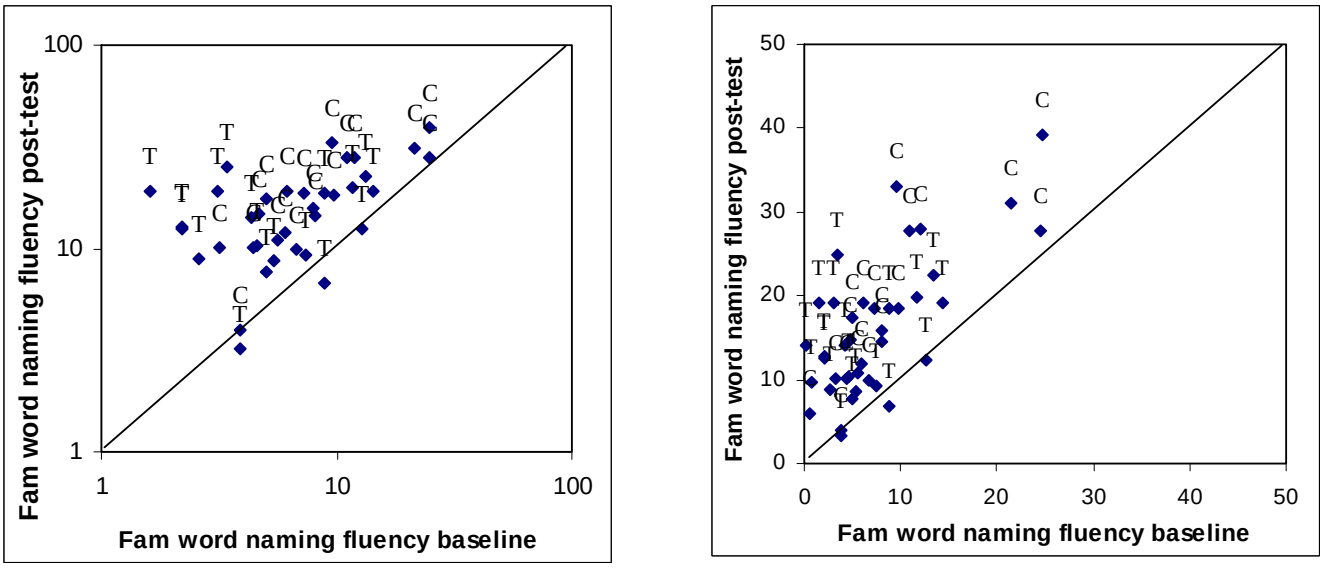
If we look at the number of students who could not perform any tasks in both treatment and control schools, and if we compare the percent change between baseline and post-intervention assessments, the analysis confirms that the treatment schools were more successful in decreasing the number of nonreaders. This was the case on all except one task, where they tied with control schools. And in the case of performance on English letter identification, control schools did better than the treatment schools.

Table 10. Percentage of students who could not read at all, compared by school type

	Control		Treatment		Percentage point improvement	
	Baseline	Post-treatment	Baseline	Post-treatment	Control	Treatment
Kiswahili letters	31%	22%	38%	16%	9	22
Kiswahili words	31%	22%	38%	25%	9	13
Kiswahili passage words	43%	25%	54%	31%	18	23
Kiswahili comprehension	82%	24%	87%	29%	58	58
English letters	23%	14%	16%	12%	9	4
English words	45%	3%	50%	5%	42	45
English passage words	47%	30%	54%	34%	17	20
English comprehension	80%	33%	85%	32%	47	53

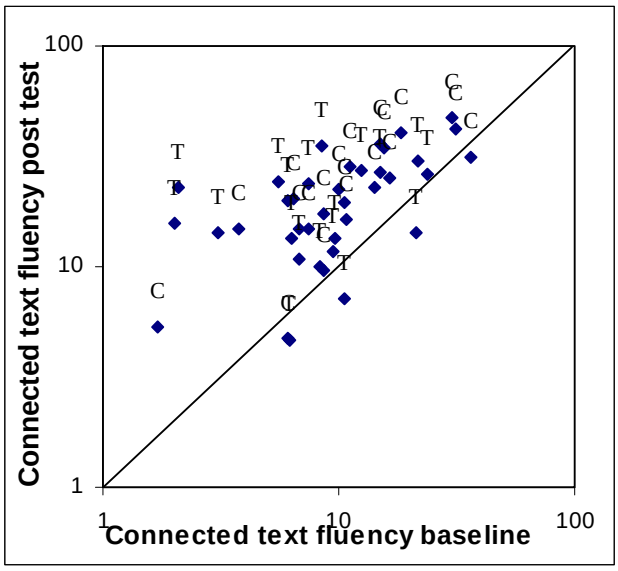
Calculated by the authors

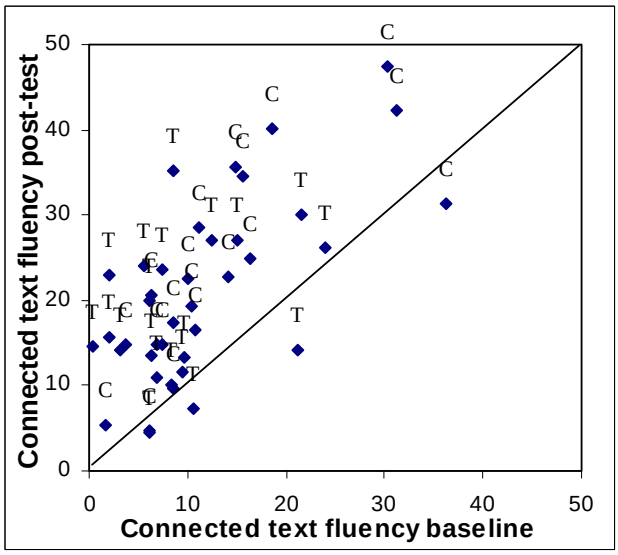
Figure 1. Pre- and post-treatment, school-level analysis



C = control; T = treatment

familiar word fluency. The substantive percentage difference (as opposed to the statistically significant difference) is very clear: *Treatment schools improved by several hundred percent*, although this might be because the treatment schools had a good few outliers that were basically not teaching reading, and were teaching reading after the treatment, so the percentage growth looms large.





C = control; T = treatment

3.2.5 A possibly important attenuating factor

As is analyzed in more depth below, it is possible that the control schools improved because of leakage of technique from the treatment schools. In some respects, if the cause of the nearly equal performance of control and treatment schools was a leakage, that could be seen as a positive thing. Unfortunately there is a less positive **possible** explanation, which is that all schools were simply improving at the same pace, for "natural" reasons that have nothing to do with the experiment. In particular, we know that even with no intervention, children's reading would improve over time. Thus, it may well be the case that the improvements observed represent a natural progression that the second graders would have made even without the intervention.

Here we explore the effect of the natural progress. As noted above, the baseline assessment took place in July 2007, while the post-treatment assessment took place in November 2008. The period between the end of July and the end of November in any given year allows for about an additional 2–3 months of learning. Our question is: Exactly how much could children have learned during those additional 2–3 months?

Life in Kenya was hardly normal during the beginning of 2008, and it is possible that teaching was not taking place in a normal environment. One could say (to fix ideas) that there were an additional two months of "normal-equivalent" instruction available to the post-test students, in a school year of nine months. But, how much would children progress normally, and does this take place at an even pace during the school year? This is not known in Kenya. However, from other developing countries, and using fluency in connected text as the key indicator, we know that the average inter-grade gain seems to be about 14 words per grade. This suggests about 1.6 words of gain per month (given nine months of instruction, 14/9=1.6).

Furthermore, from research in the United States, it is known that students typically make about 33% more progress in the first half of the year than in the second half of the year. The extra two months available to the post-test students in Kenya took place

toward the end of the school year. Thus, to adjust for this effect, and to round the numbers, one could expect a "natural" gain of about 1.35 words per month, or 2.7 words in two months, towards the second half of the school year. That is, if the gain had been due to purely natural drift or natural progression, one would have expected a gain of about 2.7 words. But the gain made in both English and Kiswahili was around nine words. So there is good reason to think that the gain went considerably beyond what would have been possible merely under a natural progression.

3.3 Need for further qualitative analysis

One is left, however, with the interesting phenomenon of approximately equivalent improvement in the control schools. The question naturally arises as to why this may have happened. Simply analyzing and comparing the baseline and post-test data cannot yield answers to this question. To answer this question it seems inevitable to try to carry out some further qualitative analysis. (It also seems imperative to prolong the intervention and carry out more quantitative evaluation, as well, during another year.) The lessons extracted also have important implications for the use of this sort of randomized study and for the nature and role of improvement in education in general. Thus, these results strike us as extremely important, and as deserving of further investigation.

The first possible explanation that comes to mind is that of intervention leakage from treatment to control schools. It may be that the control schools learned about the project, by word of mouth or some other way of information sharing, and that they simply responded to the pressure to perform better. If this were the case, then this has implications both about substantive issues related to how education improves, and also for research design. If there was some leakage, and the teachers used the improved approaches in the control schools, this suggests that, at least in Kenya, teachers are willing to improve their approaches to instruction without much external support. This is encouraging and important.

In addition to the above, it is important to determine a possible influence of district education officers on control schools. One of the aims of EGRA tools is to provide a simple diagnostic tool for teachers, and not necessarily to become a high-stakes exam. Yet Kenya is, in many respects, an exam-driven and exam-happy society. If the district officers have thought that EGRA tools bear high-stakes values, or simply enjoy assessing schools and using assessment to drive performance, then it is possible that they encouraged and provided support to the control schools in improvements of student reading levels. This needs to be looked at.

A third factor could also be a simple accountability effect or, in combination, an information effect. Schools find out how badly they are doing, and this is now expressed in a metric that is extraordinarily simple, and allows for dramatic conclusions such as "the children really can't read" as opposed to some relatively bland numerical score.

A final fourth factor is that AKF also conducted two informal assessments during the year in a (varying) sample of control and treatment schools, and it may be possible

that AKF implicitly or explicitly encouraged control schools to do better with regard to teaching reading, without meaning to, or without understanding the importance of not doing so. For instance, if the AKF officers assessed control schools and simply said "good job, but focus more on teaching reading, that is important," this statement alone has enormous impact in terms of exerting pressure on control schools that are aware they will be tested again.

4. Results of the Qualitative Research

As noted already, from the results of the post-treatment assessment done in EGRA schools in Kenya, it emerged that there was a fairly large general improvement in reading from the baseline results. However, some of the control schools had registered a large improvement too, with some doing even better than the treatment schools. This was puzzling and a qualitative follow up was necessary to understand what could have made them post such good results.

A follow-up or "forensic" qualitative analysis was therefore undertaken, with the purpose of establishing

- Whether there was leakage of the EGRA methodology and teaching materials to the control schools
- Whether there was pressure from education officers/AKF on the control schools to perform better
- What contributed to improved reading results in the control schools.

4.1 Data collection

Schools opened for the new school year on January 5, 2009. During the following week, AKF-EMACK visited nine of the control schools, purposively sampled because they had registered an improvement, plus four of the treatment schools, to find out whether they had shared some information with the control schools. The schools visited are shown below, along with their results in both baseline and post-treatment assessments.

In generating the data for the qualitative analysis, we concentrated on the control schools that improved the most: These seemed the most worthwhile candidates in which to test for strong leakage from the treatment to the control schools. In all of this, one important "leakage" vector is the possible influence of AKF's own staff on the control schools, via informal discussions and so on.

Table 11. Performance of selected schools in English[11]

School Name	Baseline School Type	Baseline Eng. Words	Post-Trtment Eng. Words	Absolute change Eng. Words	Percent change Eng. Words	Baseline Eng. Passage words	Post-Trtment Eng. Passage words	Absolute change Eng. Passage words	Percent change Eng. Passage words
SCHOOL 1	Ctrl	11	27.7	16.8	**153%**	15.7	34.5	18.8	**120%**
SCHOOL 2	Ctrl	6.2	19.1	12.9	**208%**	11.1	28.5	17.4	**157%**
SCHOOL 3	Ctrl	7.3	18.6	11.3	**155%**	10	22.6	12.6	**126%**
SCHOOL 4	Ctrl	5	17.5	12.6	**254%**	6.4	20.5	14.1	**222%**
SCHOOL 5	Ctrl	4.7	14.7	10	**213%**	8.6	17.4	8.8	**102%**
SCHOOL 6	Ctrl	4.4	10.2	5.8	**131%**	6.8	14.9	8.2	**121%**
SCHOOL 7	Ctrl	3.2	10.2	7	**217%**	3.8	14.7	11	**292%**
SCHOOL 8	Ctrl	0.6	5.9	5.3	**875%**	1.7	5.3	3.7	**223%**
SCHOOL 9	Trtment	4.3	14.2	9.9	**233%**	7.5	23.6	16.1	**215%**
SCHOOL 10	Trtment	0.2	14	13.8	**6903%**	0.3	14.5	14.3	**5708%**
SCHOOL 11	Trtment	12.7	12.4	-0.3	**-2%**	21.2	14.2	-7	**-33%**
SCHOOL 12	Trtment	5.4	8.7	3.3	**62%**	9.5	11.6	2.1	**22%**

In each of the schools visited, an interview guide was used to get information from the school head teacher and the teacher(s) who handled grade 2 the previous year. Three education officers in charge of the zones that the schools fall under were also interviewed.

4.2 Findings

4.2.1 Challenges in teaching reading

The teachers from all the schools visited highlighted some of the challenges that they faced in teaching reading, both before and during the intervention:

- Large enrollments against few teachers; the schools had an average of 65 pupils per class with one school having 120 pupils.
- Most children had a poor foundation as the preschools emphasized not letter recognition but rote learning.
- The textbooks are shared in a ratio of 1:3; hence some children do not get a chance to read them.
- The textbooks have content that is way beyond the capacity of the grade 2 child. For example, pupils are expected to read a two-page story and answer comprehension questions, yet many cannot read simple words.[12]
- Pupils prefer the use of mother tongue and avoid speaking in English.
- Pupil absenteeism makes it difficult to cover the content.

On average, the teachers said that they thought that only 55% of their pupils could read English words fluently.

[11] Schools have been made anonymous to protect their privacy.

[12] This is a phenomenon found in many countries. Not only are textbooks too ambitious relative to what children can do (sometimes grotesquely so), but so are the implicit (or even explicit) guesstimates of officials as to how well children read. As a result, teaching does not take children from where they really are to where they ought to be; it focuses on where children ought to be in a manner that is not tethered to reality.

4.2.2 Methods used in teaching reading

In each of the *control school*s visited, it was apparent that the teachers had realized, as a result of the baseline assessment, that many of their pupils simply could not read and had made great efforts to ensure that they did so by the end of the year.

The teachers employed a variety of ways to achieve this, such as the *look and say* method, recitation, and use of teaching and learning materials. Some of the teachers, such as those in School 1, School 2, and School 6, sought help from their colleagues teaching English in other classes and those at the preschool. The preschool teachers and those who had undergone a course in early childhood development (ECD) helped the teachers to learn phonics, which they effectively applied in their classes. The teachers' efforts and their commitment made a difference in these control schools.

4.2.3 Exposure to EGRA teaching methods

In two of the control schools, School 4 and School 8, each had a teacher trained on EGRA methodology posted there in the course of the year—moving from a treatment school—something very hard to control in practice. Teachers are usually transferred to a different school when promoted to the position of a deputy head teacher. As a coping mechanism, the education officers and the grade 1 EGRA teacher inducted another teacher in the school to teach the grade 2 class. The two teachers in School 4 and School 8 said that they used the EGRA methodology in their new schools as the "reading levels were very low." This could explain the improved performance (School 4 with 254% improvement, and School 8 with 875% improvement) in the two schools.

In School 5, the head teacher was instrumental in finding out how to improve reading. This was after he discovered that his son, who was in grade 1 in a neighboring treatment school (School 9), could read after only a few months in school. He said that he inquired from the Education Office on why his school was not implementing the EGRA methodology and was told that this was an experiment and his school was a control. He was not happy with that and he decided to learn the methods. He sent his lower primary school teachers to find out what "secret methods" the teachers were using. One of the teachers was also proactive when she saw a teacher who is her neighbor and works at School 10 (a treatment school) making lots of teaching aids. She said:

"I asked her why she was always making flashcards, word charts and puzzles. She told me that they helped her teach reading. I decided I had to do the same for my class" —Grade 2 teacher, School 5

In the course of discussions, the head teacher said that he wanted to fully implement the program in the school this year and requested EMACK's help. He said that if this could be provided, he was ready to organize a workshop in the school to be facilitated by the EGRA teachers from neighboring schools. This almost subversive (in the sense of a militant attitude toward self-treatment) but flattering appreciation of EGRA was quite amazing (certainly something we did not expect), and suggests interesting methodological considerations in doing controlled studies when schools are near each

other (but if they are not, then they are imperfect controls in other respects—a conundrum).

Leakage could also be detected by visiting the treatment schools. It was found that in two of the schools (School 9 and School 10) there was leakage of the EGRA materials with neighboring control schools due to close proximity and interactions of the teachers. The teachers said that they used to make the teaching and learning materials at home and since they share the same compound with teachers in the control schools, their colleagues would be curious and want to know what they were doing.

The issue of informal or implicit (or perhaps merely possible) leakage was also interesting. At School 11 and School 12, there was no leakage of the EGRA methods and teaching materials to the control schools. The teachers interviewed, however, said that teachers from the neighboring schools were curious about what method they were using to enable children in grade 1 to read.

"I have three children in my class whose parents teach in the neighboring schools (not part of the treatment group). They have all asked me at different forums what methods I was using to achieve that. I only told them I make use of teaching and learning materials. However none of them have visited me in school." —Grade 1 teacher at School 12

In some of the treatment schools visited, it was found that the grade 1 children who had been exposed to EGRA were better readers than the grade 2 pupils who had also undertaken EGRA lessons. The reasons for this were not clear, and should be explored further. It is possible that the EGRA experience should be extended to grade 1.

The qualitative or "forensic" research also revealed another phenomenon: The teachers quickly picked up on the notion that this highly directed approach could help children in later grades who were having reading difficulties. Of course, no quantitative evidence is available on either the degree to which this actually happened, or whether these children were able to improve. It is nonetheless interesting.

Another interesting issue arose during analysis and reflection based on this "forensic" research. Because some of the control schools were selected to be a little distant from the treatment schools, precisely because there was intuition that there might be contamination (and there was contamination anyway, as this qualitative research demonstrates), then the control schools are not the same as the treatment schools other than for the fact of treatment. That is, other things being equal *(ceteris paribus)*, conditions did not exactly hold.

Because the treatment schools tended to be more from within the EMACK II project (treatment required access and prior infrastructure, and this was obviously only possible by using project schools), they tended to be in the inner portion of Malindi District, whereas the control schools tended to be more from the peri-urban coastal zone. It is possible that the control schools were therefore more "aware," or "progressive" in general, and more likely to be affected simply by the attention and

measurement. We do know that in *some* of the treatment schools teachers thought that the EGRA innovation was "too much work" whereas in *some* of the "contaminated" control schools the teachers were actually eager to try the method.

One clear lesson is that these issues have to be very carefully considered in any further experimentation of this nature. Preventing contamination by physical separation, however, as noted elsewhere in this paper, does introduce other biases, though, because the schools by definition will not be the same as the treatment schools if purposefully physically separated. If the groups are separated by having the randomization operate over such a large area that all the schools are far from each other, however, one is then creating an experiment that in some sense is not faithful to itself or to the eventual likely intervention. This is the case because eventual implementation is not done over large areas, and the intervention being modeled is not just what the teachers do, but the delivery of help to the teachers.

4.2.4 Was subtle or direct pressure exerted to teach reading?

It is possible that the control schools faced a sort of accountability pressure, and, certainly, as noted above, even the awareness (an information function) created in the control schools by the mere baseline seems to have awakened a sense of urgency in teachers. (This is one of the advantages of using a very direct and simple measure of outcomes.) In discussions with the teachers and head teachers, it was evident that the desire to compete exerted quite some pressure on the teachers. The teachers, even in the control schools, knew that they were being assessed together with other schools and did not want to look bad. From the assessments done by EMACK and the District Education Office in May and August 2009, the teachers would get immediate feedback. Remarks such as "your children cannot read" or "you have good readers" motivated the teachers to put more emphasis on reading.

"After the assessment there was someone who told me the children can read better if they connect words in a sentence. So I started making them recite words, using flash cards and encouraging them to speak in English. I also assigned more time to oral work." —Grade 2 teacher at School 2

In School 6, the deputy head teacher was unhappy with the performance of his school in the first assessment. He then took a personal interest in EGRA as he is a trained ECD teacher. With his strong ECD background, he helped the grade 2 teacher improve the teaching of phonics and helped the teacher use more teaching and learning materials, and use them better.

It was found that the EMACK staff and education officers did not directly exert pressure on the control schools to perform well. However, the feedback given after each assessment served as a catalyst or implicit pressure to work hard in the schools visited.

4.3 Conclusion on qualitative analysis

From the above analysis, it does appear that some of the control schools were "leaked" techniques. It also appears that in the control schools there was an implicit or explicit informational or accountability pressure, or simply awareness-raising. This is always a possibility in a social research initiative as people are always curious and interact freely. In essence, the lack of a placebo (as is used in medical science) or of alternative treatments, and the ease of copying the treatment and hence self-treating, did result in what appears to be considerable self-treating, both by using EGRA tools and simply by making more effort in general, in the control schools. It should be noted that some of the schools were within a 4 km radius from each other and many teachers lived in the same town and neighborhoods. There is also the "density" of the experiment: 40 schools (20 treatment and 20 control) in a district with only 120 schools. This can increase the likelihood of leakage. We come back to this issue below.

It would appear sensible to recommend, in future, that one should have the control and experimental schools in different districts but with similar socioeconomic status. However, in some sense, picking schools that are far from each other would introduce other unobservable factors, and in some ways would violate the purpose of randomization. In effect, one would be violating the condition of "all other things being equal." There is no obvious solution to the problem, except perhaps to provide two treatments, to see which one is more effective, and thus "fool" the control group; or have the two treatments serve as controls for each other. An alternative, being tried by RTI in Liberia at the suggestion of the World Bank, is to provide one treatment that explicitly uses accountability and informational pressure (as a "lighter" treatment), and nothing else, and a "fuller" treatment that uses those two factors and training and teacher development as well, with a control group of schools that are never visited for at all for improvement reasons and are not told they will be retested.

Most of the teachers were keen to learn a new method that would make a difference in the children's reading ability. Some of those in the control schools went out of their way to copy the EGRA teaching methods.

5. Overall Conclusions: Is the Intervention Working?

Results clearly improved. The qualitative research strongly suggests not only that results improved, but that uptake by teachers was quite positive, with techniques rather unavoidably leaking from treatment to control schools. Thus, fluency in reading, for example, improved markedly. Fluency in Kiswahili letter-naming, a skill found to be particularly poor in the baseline, improved remarkably. If something is taught, and taught well, directly, explicitly, it is learned.

It needs to be noted, however, that the intervention did not achieve the nominal goal of reaching, say, 45 correct words per minute (agreed upon during the initial stakeholder design workshop in April 2007) in connected-text fluency. In that sense,

there is still a lot to be done, and the intervention has not worked nearly as well as one might have hoped. Yet, there does seem to have been considerable impact. Furthermore, the qualitative analysis reveals that schools and administrators, locally, can in fact become excited about the intervention, since the intervention spontaneously leaked to other schools. RTI's experience in other countries suggests that local implementers (government or nongovernment) are excited by innovations, but often do not realize how difficult it is to get to a given goal. It is also possible, as many hypothesize, that merely providing information, and either explicit or implicit accountability pressure, spurs schools to do better. There were strong suggestions of this in the qualitative research.

Making a change at the school level requires a lot of hard work and dedication to changing teacher behaviors. In South Africa, through USAID's flagship projects (District Development Support Program [DDSP], and its successor, the Integrated Education Program [IEP]), RTI eventually saw very significant improvements. But RTI tended to see no improvements whatsoever in the first two years of the project, even after significant investments in teacher training and resources. The approach, it turned out, was too generic, and too broad—too much about overall management of the schools, and insufficiently focused on results. What was lacking, then, was a systematic approach not only to teaching reading, but also—and more importantly—to supporting teachers on ongoing basis. Resources such as decodable books, parental commitment to ensuring that children read at home at least 20 minutes a day, and other teacher resources are an important aspect of the support to teachers.[13] The realization of just how intense and focused the intervention has to be takes some time to set in, and is helped by evidence of the kind reported in this document. The support has to be systemized, uninterrupted, intense, and long enough to ensure that the teachers do receive adequate support. The instruction in reading has to be direct and explicit, has to follow a specific scope and sequence, and has to contain clear lessons that fit within that scope and sequence. It is likely that the reading intervention in Malindi needs to be intensified in these respects.

Further analysis should be done (if possible, either with existing data or with further data generated) on the precise impact of the training per se. Emphasis thus far has been on the lesson plans as an intervention, with insufficient research on the training of the teachers. AKF and RTI should further jointly think about what kind of training and resource materials were provided to teachers as well as how the support visits to schools were structured. This, in addition to the above suggested follow-up discussions with teachers, would shed more light on possible lack of fidelity in the intervention.

Alternatively, of course, it may be necessary to "simply" implement with more rigor. Fidelity checks would help determine how to intensify and would by themselves increase the rigor. More effort simply tracking or assessing the use of time might also be helpful, since the "intervention dosage" is an important consideration, and really perhaps not sufficiently evaluated in this case. Of course, in these cases the

[13] By "decodable books" we mean books containing only words with no irregular print-to-sound correspondence.

intervention is more expensive, and sustainability questions arise. However, if the purpose is to assess the utility of the intervention as such, during an experimental or testing phase, fidelity is naturally an important consideration. And if "dosage" (time used for teaching reading) is an important consideration, but the government is serious about getting reading taught well and early, then sustainability would require that time-tabling and time reserved for teaching reading be made an explicit matter of government policy as well as ongoing supervision. Kenya, perhaps somewhat differently from other developing countries, does have a well-developed school support system at the zonal level. Strategies that show serious improvement results can be generalized.

It would be a missed opportunity not to further support AKF in an attempt to more exactly pinpoint reading improvements. RTI would stand ready to collaborate and help organize another evaluation in November 2009. Another advantage of being able to assess again in 2009 would also be that the issue of whether there is simply natural progression, or real impact, would be addressed simply by testing the children at exactly the same time period, a desirable thing that was not possible due to the slowness in start-up.

Methodologically, it seems wise to continue, and perhaps to become cleverer about the use of more than one treatment or the use of further-away control schools. On the other hand, one needs to be cautious about attempting too many elaborations on a basic methodology, if the evidence of the experiment suggests to teachers and principals themselves that the treatment is working, and they demand it, to the point of self-treating. A good strategy, from both ethics and research design points of view, would be to roll out the intervention to the ex control schools, and add another control group, so as to have a sort of rolling randomization that continues to help one assess intervention impact, while extending intervention.

It also seems wise to extend the intervention more officially to grade 1. Teachers in grade 1 were being trained, but the evaluation design did not evaluate the impact on grade 1. This does raise issues if further intervention and evaluation is done, since some grade 2 children will have had two years of exposure to the techniques. The uptake in grade 1 may have been even better—though it was not tracked. This could also form the basis of a longitudinal study, if done properly.

It would be important to investigate how teacher characteristics such as gender, age, job satisfaction, and qualification affect their implementation of EGRA. Besides assessing children's reading baseline, a baseline for teachers' professional competency in the teaching of the English language should be undertaken. This would provide a solid basis for eliminating various biases or limitations in the research initiative.

Finally, it is useful to draw out some of the methodological implications for evaluation. The gold standard is, of course, randomized evaluation, ideally with pre- and post-assessment. However, when there is no placebo, and when the treatment is relatively easy to copy (so that the controls can self-treat) once extended to the treatment schools, there are all sorts of interesting issues that come up.

There is of course the leakage issue. No alternative that we can think of truly deals with this problem in a satisfying way. Putting the control schools further away would violate the very purpose of randomization as a tool for generating *ceteris paribus* conditions; putting the control schools further away, as a group, would mean that those schools were clearly different in some way. One could disperse the control schools over the whole country, so that they were not clustered in some other presumably similar (but subject to unobserved third variable influences). But the treatment schools would be clustered and thus possibly different. One could, of course, disperse both the control and treatment schools widely over the country, or, at any rate, so widely that they could not influence each other.

But one needs to think carefully about this. After all, the treatment one is testing here is not, as in so many experiments, a very simple thing, certainly not as simple as giving someone a pill. One is, in a sense, testing a treatment (the lessons), but also the delivery vector (the training, the supervision). And in fact one, in some sense, wants to encourage the treatment schools to talk to each other, but not to talk to the control schools. Scattering the schools widely abstracts away from key features of the treatment and hence adds a sort of ersatz (or perhaps real enough) rigor, *but* at the cost of forcing an evaluation of something that is inherently more trivial (and thus perhaps less powerful, and certainly less interesting in its implementation or scale-up potential) than the "real" intervention one seeks, not to mention the huge costs it would involve.

So, the attempt to be rigorous actually forces a redefinition of what is being evaluated and what is evaluable—an interesting scientific conundrum. It is not an accident that so many of the assigned randomized control experiments (as opposed to the natural experiments or the regression discontinuity exercises) often relate to interventions that seem in some sense fairly trivial. (Perhaps it is true, however, that there is an implicit philosophical point here: better trivial but well tested than profound but difficult or impossible to test.) Other designs (such as regression discontinuity or randomized inclusion in a non-experimental program) that take advantage of natural randomization do not suffer from these problems, but are not suitable for this kind of reading intervention, or none have been found yet.

Finally, the fact that interventions seem to have leaked to the control schools suggest that, at least in countries such as Kenya, where teachers seem reasonably well motivated and there is a tradition of accountability and responsiveness to measurement, suggests that teachers are in fact quite willing to take up innovations, if they are straightforward and seem to show results. (Note, however, that some of the teachers deemed the intervention to be too much work, and probably did not implement with fidelity or did not implement much at all.)