Policy Research Working Paper 8427

A Randomized Evaluation of a Low-Cost and Highly Scripted Teaching Method to Improve Basic Early Grade Reading Skills in Papua New Guinea

Kevin Macdonald
Binh Thanh Vu1

WORLD BANK GROUP
Education Global Practice
May 2018

Policy Research Working Paper 8427

Abstract

Early grade literacy skills are crucial for children's future education and ultimately their contribution to human capital formation and economic development. A significant challenge in development is identifying low-cost interventions to improve early literacy skills in contexts characterized by varying teacher ability and severe budget constraints. This paper evaluates the impact of Papua New Guinea's randomized Reading Booster Programme, which was conducted in Madang and Western Highlands Province in 2013 and 2014, respectively. The program provided teachers with training on a highly structured teaching method that they could apply one hour per day within the teaching time allocated to reading. Using the randomized assignment of schools into the program, the paper shows that it had a substantial impact on the reading skills targeted by the program for third grade students, ranging from 0.6 to 0.7 standard deviation. Large effects on other reading skills were found for girls but not boys. The program's cost per student was approximately US$60.

This paper is a product of the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at tvu@worldbank.org. and kadmacdonald@gmail.com.

Produced by the Research Support Team

A Randomized Evaluation of a Low-Cost and Highly Scripted Teaching Method to Improve Basic Early Grade Reading Skills in Papua New Guinea

Kevin Macdonald[1]
Binh Thanh Vu[1]

Keywords: Human capital, randomized-controlled trial, literacy, primary education, cost effectiveness, Papua New Guinea

JEL classifications: I24, I25, I28, O15

[1] World Bank. The views expressed by the authors may not reflect that of the World Bank Group. The authors are grateful to Felipe Barrera, Chris Thomas and Myrna Machuca-Sierra for their comments. Special thanks to guidance and comments provided by Harry Patrinos. The authors may be contacted at kadmacdonald@gmail.com and tvu@worldbank.org.

1. Introduction

The relationship between young children's literacy skills, future human capital formation, and subsequent economic development is an increasingly researched theme in economic development. The importance of literacy to individual productivity including the diffusion of technology is well established in developing countries (Basu and Foster 1998; Rosenzweig 1995), and literacy has been described as a threshold for economic development (Azariadis and Drazen 1990). Gaps in reading skills persist as children age (Butler et al. 1985); as a result, early literacy skills are an important determinant of a child's future education outcomes including future literacy skills (Marteleto et al. 2008; Entwisle et al. 2005; Jimerson et al. 2000; Alexander et al. 1997).

Despite their importance for development, developing countries struggle to provide children with basic literacy skills, even after substantial progress towards the 1990 Education for All goals. For example, in a recent regional assessment of 10 countries in Francopone Africa, the *Programme d'analyse des systèmes éducatifs de la confemen* (PASEC), 71.4 percent of 2nd grade students and 57.3 percent of 5th grade students on average do not achieve minimum proficiency in literacy (PASEC 2015:36,50). In the Pacific, assessments of early age literacy in Tonga in 2009 and in Vanuatu in 2011 found that after three years of schooling, only 30 percent of students in Tonga and 25 percent of students in Vanuatu were able to read fluently for comprehension (World Bank 2012a, 2012b, 2012c). In Kiribati and Tuvalu, 20 percent of 3rd grade children achieved minimum reading comprehension proficiency (World Bank 2017a, b). In Papua New Guinea, early grade reading assessments conducted in four provinces between 2011 and 2013 found that students lagged two years behind curriculum targets for fundamental pre-reading skills (World Bank 2014a, 2014b, 2014c, 2014d).

International research has identified basic skills that young children need in order to read alphabetic languages (Linan-Thompson and Vaughn 2007; Wolf 2007; Sprenger and Charolles 2004; Chiappe et al. 2002; see also: Gove and Cvelich 2011 and National Reading Panel 2000). Among these are an understanding of the relationship between printed letters and sounds (Scarborough 2002), the speed at which a child can read (Abadzi 2006), and oral reading fluency (Fuchs et al., 2001).

An important challenge is how to ensure children in the early grades of school acquire these skills in a context where teachers have varying and often few formal qualifications, implementation capacity is weak, and budgets allow for no or limited expenditure beyond teachers' salaries. In this context, several randomized controlled trials of interventions involving heavily scripted and systematic instructional approaches have been shown to be successful at improving early grade reading outcomes. The South Africa Integrated Education Programme implemented the Systematic Method for Reading Success improving early grade reading skills by approximately 0.8 standard deviations (Piper 2008). The EGRA plus program in Liberia demonstrated impacts of 0.52 to 1.23 standard deviations (Piper and Korda 2010), and the Mali Read Learn Lead program improved reading outcomes by 0.3 to 0.6 standard deviations (Spratt, King and Bulat 2013).

This paper evaluates a similar though smaller scale approach of scripted and systematic instruction piloted as a randomized controlled trial in two provinces in Papua New Guinea.[2] In this approach, teachers use one hour-long lesson per day from the time allocated to language instruction to follow highly scripted lessons on teaching reading skills. Unlike other approaches, this is not a comprehensive reading instruction program, but rather provides remedial lessons on reading that is low cost to implement. As a result, this paper contributes to the literature by showing that low-cost, scripted instructional approaches in a remedial course format can have significant effects on early grade reading skills in a developing country context.

2. Papua New Guinea and the Reader Booster Programme

Papua New Guinea is classified by the World Bank as a lower-middle-income country. Its per capita gross national income was 2,240 USD in 2014. Economic growth has been around 5 percent per annum over the last 16 years, and an estimate of 6.6 percent per year between 2012 and 2016. The projected population in 2015 is 7.6 million people, and the adult literacy rate is projected to be 63.4 in 2015 (World Bank 2016a).

[2] This program was implemented by the Government of Papua New Guinea under the Reading Project in PNG (P105897) funded by the Global Partnership for Education (GPE) with technical support from the World Bank and Marcia Davison.

Its primary education sector has experienced significant growth in participation: from 2008 to 2012, during which time its universal basic implementation plan was implemented, primary enrollment increased from 600,000 to 1.4 million students. The primary gross enrollment ratio increased from 60.5 to 114.7, and the primary net enrollment rate was 86 percent according to the latest available data from 2012 (World Bank 2016b). Children begin school with three years of elementary school followed by six years of primary school.[3] The language of instruction in elementary school is a local language while English is the official language of instruction for primary schools.

Despite significant increase in enrollment, learning outcomes remain poor. Eighth grade exam results reveal poor outcomes for literacy and numeracy (World Bank 2011). Early grade reading assessments conducted in the National Capital District, Madang Province, Western Highlands Province and East New Britain from 2011 to 2013 found that children's crucial pre-reading skills including "alphabetic principle and phonetics" were two years behind the curriculum target. They also found that students took five years to attain some reading skill objectives required by the first-grade curriculum (World Bank 2014a, 2014b, 2014c, 2014d).

The Reader Booster Programme was designed as a remedial course to improve early literacy skills for children in the first two years of primary school, grades 3 and 4 (see Government of Papua New Guinea 2015; 2016). Kindergarten and grades 1 and 2 are typically taught at elementary schools. The program was implemented in Madang Province in 2013 and Western Highlands Province in 2014. The intervention developed highly scripted lesson plans for teachers to follow for one lesson per week and provided teachers with training to implement follow-up lesson plans. Time for these lessons was scheduled during the curriculum time allocated for language instruction; the national curriculum allows teachers to have flexibility in which materials they use and how they teach. In addition to training, teachers were also provided with mentoring and coaching.

[3] Until 2017, the general education system in PNG includes 12 years of schooling divided by elementary schools offering a three-year program consisting of a preparatory year and grades 1 and 2, and primary schools offering a six-year program consisting of grades 3-8.

The intervention targeted three key pre-reading skills: initial sound identification, letter sound knowledge, and word reading. These domains in addition to several others were tested in the series of early grade reading assessments conducted before and after the interventions were implemented.

In both provinces, schools were randomly sampled and assigned to either a treatment group which received the intervention or a control group which did not. In Madang Province, 15 schools were assigned to the treatment group while 16 schools were assigned to a control group. In Western Highlands Province, 23 schools were assigned to the treatment group and 23 to the control group.

The intervention was implemented in the treatment schools in Madang province in 2013. However, the unpublished government report indicates that the intervention was delayed until late in the school year due to various logistical issues; consequently, the Madang students were not fully exposed to the intervention (Government of Papua New Guinea 2016). The intervention in the Western Highlands Province was implemented in 2014.

3. Data and sample characteristics

A. Early grade reading assessment and timeline

In order to measure the impacts of the interventions, an Early Grade Reading Assessment (World Bank 2014b,c) was applied the year before and the year after each intervention in Madang and Western Highlands Provinces. The assessment measures basic reading skill domains: letter recognition, phonemic awareness, phonics, word reading, oral reading fluency, reading comprehension, listening comprehension and alphabetic principle. These skills are measured on nine sub-tests: letter name knowledge, initial sounds of words, letter sound knowledge, familiar word reading, unfamiliar word reading, reading comprehension, listening comprehension, oral reading fluency and dictation (World Bank 2014a:23). The intervention aims to improve three reading skill domains measured in these data: word initial sound knowledge, letter sound knowledge and familiar word knowledge. However, the letter sound domain is excluded from this analysis as the government report as well as an unpublished reliability analysis found the domain to be unreliable in the end-line Madang dataset.[4]

The Early Grade Reading Assessment was conducted four times, both before and after the interventions in the two provinces. In Madang, the intervention occurred in 2013, and the assessments were conducted in 2011 and at the end of the school year in 2013. In Western Highlands Province, the intervention occurred in 2014, and the assessments were conducted in 2013 and at the end of the school year in 2014. The data sets were not implemented as a panel as a new sample of students was drawn in each round.

The World Bank provided four data sets for each of the four rounds of the assessment. These data included scores for each of the 9 reading skill domains, sample weights, and several variables about the schools and students, which are described below. The reading domain scores contain a proportion of zero scores, which vary depending on the domain, suggesting a truncated distribution. As a result, the reading domain scores are standardized using a mean and standard deviation of the control group, baseline students estimated using a Tobit model.

B. Sample sizes and attrition

Table 1 presents the number of schools sampled in each round of EGRA. In Madang province, the baseline sample included 11 of the 15 control schools and 10 of the 16 treatment schools. End-line assessment data are missing for 5 of the sample control schools and 3 of the sampled treatment schools. The unpublished government report states that this is due to logistical, financial and weather issues; it is unlikely that the interventions had any effect on school attrition in Madang. Four of the remaining control schools and 5 of the remaining treatment schools were added to the sample, but there are no baseline data for these additional schools. In Western Highlands Province (WHP), 10 each of the 23 treatment and control schools were sampled at baseline and all 23 treatment and control schools were sampled at end-line. For the additional 13 control and 13 treatment schools included in the end-line data, there are no baseline data.

[4] An unpublished reliability analysis found that the letter sounds domain was negatively associated with the initial sounds domain controlling for all other domains for the Madang end-line sample. For other samples, it is strongly positive. We replicated this analysis and found the same result. If it were included in this paper, it would yield the largest effect size (over 1 standard deviation) compared to the effect sizes on the other domains.

Table 1. Trial and data summary

Madang Province	Control	Treatment
No. of schools in the trial	15	16
of which were sampled in EGRA		
at baseline (2011)	11	10
at end-line (2013)	10	12
at both baseline and end-line	6	7
Western Highlands Province	**Control**	**Treatment**
No. of schools in the trial	23	23
of which were sampled in EGRA		
at baseline (2013)	10	10
at end-line (2014)	23	23
at both baseline and end-line	10	10

Source: PNG Madang and WHP EGRA datasets for number of schools in the sample; unpublished government reports and sampling data for number of schools in the pilots

Sample sizes by grade, province and school treatment status are described in Table 2. Second grade students were sampled only in the Madang province baseline round, and the samples for the Western Highlands Province include 4th grade students only at baseline. Each round of EGRA is sampled as a repeated cross-section, and, because of the timing, no cohort was sampled more than once except for the 2nd grade students in the Madang baseline; they were in 4th grade at the time of the Madang end-line sampling.

Table 2. Number of students sampled in EGRA

	Madang Province		Western Highlands Province	
	Control	Treatment	Control	Treatment
Baseline sample				
No. of 2nd grade students	188	212	0	0
No. of 3rd grade students	230	216	223	228
No. of 4th grade students	214	219	211	227
End-line sample				
No. of 2nd grade students	0	0	0	0
No. of 3rd grade students	197	240	466	489
No. of 4th grade students	184	254	0	0

Baseline and end-line surveys sampled different cohorts of students. Source: PNG EGRA Madang and WHP datasets

The data used in this evaluation are that of grade 3 only. The 4th grade sample at end-line does not include any students from Western Highlands Province, and while the 4th grade sample in Madang province could be used with the 2nd grade sample, the sample of students in schools that were included in both baseline and end-line is small.

Table 3 compares baseline reading achievement scores between students in schools that appeared in the baseline and not the end-line ("attrition schools") and in schools that appeared in both the baseline and end-line ("non-attrition schools"). None of the reading domains' differences are statistically significant; however, a difference as large as 0.3 standard deviation cannot be rejected for the letter names domain. With the exception of this domain, the data suggest little difference in reading achievement between the attrition schools and schools appearing in both rounds of the survey.

Table 3. Baseline difference in reading scores between attrition schools and schools sampled in both base- and end-lines (standard deviations).

	Attrition schools	Non-attrition schools	Difference
Reading domains targeted by the intervention			
Initial sounds	0.143	0.286	-0.142
	(0.101)	(0.039)	(0.108)
Familiar words	0.208	0.167	0.041
	(0.183)	(0.058)	(0.192)
Other reading domains			
Letter names	-0.097	0.092	-0.189
	(0.245)	(0.043)	(0.249)
Unfamiliar words	0.08	0.169	-0.088
	(0.15)	(0.046)	(0.157)
Reading comprehension	0.166	0.252	-0.086
	(0.103)	(0.046)	(0.113)
Oral comprehension	0.08	0.179	-0.099
	(0.113)	(0.05)	(0.124)
Dictation	0.136	0.172	-0.036
	(0.26)	(0.058)	(0.266)
Oral reading fluency	0.107	0.188	-0.08
	(0.151)	(0.059)	(0.162)

Standard errors presented in parentheses. Statistical significance at the 10, 5 and 1 percent levels denoted by *, **, and ***, respectively.

Students in attrition and non-attrition schools differ in terms of their background characteristics, as compared in Table 4, but neither has a clear advantage. Students in attrition schools are less likely to have printed materials at home to read and more likely to be absent for more than two weeks from school in the previous year but, at the same time, are more likely to have someone read to them at home and are in smaller classes. They are also less likely to be in multi-grade classes and be tested in the national language, *Tok Pisin,* rather than English. Neither attrition nor non-attrition school students have a consistent advantage in terms of background characteristics.

Table 4. Baseline difference in background variables between attrition schools and schools sampled in both base- and end-lines (standard deviations).

	Attrition schools	Non-attrition schools	Difference
Female student	0.524 (0.026)	0.477 (0.01)	0.047 (0.028)
Speaks English at home	0.39 (0.047)	0.385 (0.024)	0.005 (0.053)
Has printed materials at home to read	0.629 (0.03)	0.696 (0.022)	-0.067* (0.037)
Has someone at home who can read	0.879 (0.024)	0.853 (0.018)	0.026 (0.03)
Someone reads to the student at home	0.854 (0.027)	0.674 (0.026)	0.18*** (0.037)
Absent more than two weeks in prev year	0.49 (0.067)	0.245 (0.023)	0.245*** (0.071)
Multigrade class	0.14 (0.128)	0.919 (0.032)	-0.779*** (0.132)
Class size	41.371 (2.678)	78.188 (10.888)	-36.817*** (11.213)
Majority of EGRA tests in Tok Pisin	0.004 (0.004)	0.196 (0.048)	-0.192*** (0.048)

Table presents estimates of the difference in differences in background variables at baseline between students in treatment and control schools and in schools included and not included in the end-line sample. Standard errors presented in parentheses. Statistical significance at the 10, 5 and 1 percent levels denoted by *, **, and ***, respectively.

C. Baseline balance

Table 5 presents the difference in baseline reading achievement scores between treatment and control group students for those in all schools sampled at baseline and for those in schools appearing in both the baseline and end-line samples; positive values indicate that the treatment group has a higher estimate than the control group. For all baseline schools, large and statistically significant differences exist between the treatment group and control in three reading domains: familiar words, dictation and oral reading fluency. Several other domains have differences that, while not statistically different from zero, are also not statistically different from 0.2 standard deviation. In other words, moderate differences between the treatment and control groups cannot be ruled out. In the sample of students in non-attrition schools, the differences between treatment and control groups tend to be lower. None of the reading domain differences for these students are statistically different from zero, but several, including dictation and oral reading fluency, are not statistically different from 0.3 standard deviation as well.

Table 5. Baseline differences in reading scores between treatment and control groups at baseline (standard deviations)

	All baseline schools	Non-attrition schools
Reading domains targeted by the intervention		
Initial sounds	0.091	0.046
	(0.074)	(0.078)
Familiar words	0.246*	0.169
	(0.14)	(0.161)
Other reading domains		
Letter names	0.122	0.041
	(0.094)	(0.099)
Unfamiliar words	0.132	0.058
	(0.099)	(0.112)
Reading comprehension	0.112	0.06
	(0.106)	(0.134)
Oral comprehension	0.036	0.014
	(0.104)	(0.123)
Dictation	0.288**	0.159
	(0.14)	(0.154)
Oral reading fluency	0.251*	0.179
	(0.143)	(0.17)

Table presents estimates of the difference in reading score between treatment and control groups. Positive differences imply that the treatment group has a higher score than the control group. Standard errors presented in parentheses. Statistical significance at the 10, 5 and 1 percent levels denoted by *, **, and ***, respectively.

Differences in the available background variables between treatment and control groups are estimated in Table 6. Statistically significant differences exist for the proportion of females, availability of printed materials at home, whether someone reads to the child at home, class size, and whether a majority of the tests at the school are in *Tok Pisin*. The differences between treatment and control groups are roughly the same whether comparing students in the baseline schools or students in the non-attrition schools. This suggests that any imbalance is a result of the randomized assignment of treatment rather than the attrition of schools.

Table 6. Differences in background characteristics between treatment and control groups at baseline

	All baseline schools	Non-attrition schools
Female student	-0.051***	-0.055***
	(0.016)	(0.018)
Speaks English at home	0.07	0.085
	(0.051)	(0.058)
Has printed materials at home to read	0.085**	0.088**
	(0.037)	(0.042)
Has someone at home who can read	0.038	0.025
	(0.034)	(0.04)
Someone reads to the student at home	0.093*	0.107*
	(0.054)	(0.057)
Absent more than two weeks in prev year	-0.039	-0.063
	(0.052)	(0.064)
Multigrade class	-0.004	-0.012
	(0.092)	(0.079)
Class size	40.822**	46.318**
	(18.425)	(21.544)
Majority of EGRA tests in Tok Pisin	-0.174*	-0.163*
	(0.091)	(0.093)

Table presents estimates of the difference in background variables between treatment and control groups. Positive differences imply that the treatment group has a higher value than the control group. Standard errors presented in parentheses. Statistical significance at the 10, 5 and 1 percent levels denoted by *, **, and ***, respectively.

4. Empirical strategy and results

A. Estimation model

The empirical strategy is to estimate the school-level effect of the reader boost program using a difference-in-differences approach with covariates. The school-level effect is estimated because students in the baseline and end-line samples are different and represent different cohorts. The difference-in-differences approach and the inclusion of student and school background variables as controls are motivated by the imbalance detected between treatment and control groups in some achievement scores and background variables. The effect is also assumed to vary by gender.

Reading scores for the i^{th} student and school j, Y_{ij} are modeled as a linear function of being in a program school, P_j, being sampled at end-line, t_{ij}, being female, F_{ij}, other student and school characteristics, $\boldsymbol{X}_{ij}$ and disturbance, u_{ij}.

$$Y_{ij} = \beta_0 + \beta_1 P_j + \beta_2 t_{ij} + \beta_3 t_{ij} P_j + \beta_4 F_{ij} + \beta_5 F_{ij} t_{ij} + \beta_6 F_{ij} P_j + \beta_7 F_{ij} t_{ij} P_j + \boldsymbol{\beta_8} \boldsymbol{X}_{ij} + u_{ij}$$

Coefficient, β_3, is the impact of the program on male test scores, and $\beta_3 + \beta_7$ is the impact on female test scores. Because achievement scores in some domains may have a high proportion of zero scores, a Tobit model is used to estimate the model. Baseline sample weights are adjusted to reflect attrition of schools in the end-line data, and standard errors are estimated to be robust to the two-stage sampling method (schools, then students) and a finite population correction based on the number of schools in each province.

B. Impact of the Reader Booster Programme

Estimates of the model are presented in Table 7. For males, the intervention has a statistically significant and large effect on one of the reading domains targeted by the intervention, of 0.63 standard deviation for initial sounds. The effect on males' familiar words achievement is not statistically different from zero. For females, the effect is large and statistically significant for both domains, and statistically higher than males for the familiar words domains.

The program also had positive effects on reading domains that are not targeted by the intervention. For males, a positive effect is found only for oral reading comprehension; for females, positive and large effect sizes are found for five of the six other domains. The effect size in the other domains is statistically higher for females than males in three of the six domains.

Table 7. Tobit model estimates of the impact of interventions on reading domains (in standard deviations). Includes schools in both baseline and end-line surveys only

	Domains targeted by the intervention		Other reading domains					
	Initial sounds	Familiar words	Letter names	Unfamiliar words	Reading comp.	Oral comp.	Dictation	Oral read. fluency
time	0.027	0.562***	0.529***	0.42**	0.412*	-0.178	0.335	0.695***
	(0.178)	(0.199)	(0.153)	(0.177)	(0.22)	(0.161)	(0.204)	(0.215)
treatment	0.103	0.044	0.072	0.113	0.091	-0.164	0.068	0.1
	(0.129)	(0.127)	(0.115)	(0.16)	(0.224)	(0.146)	(0.146)	(0.137)
time x treatment	0.627***	0.304	-0.03	0.121	0.245	0.485*	0.063	0.12
(**impact on males**)	(0.186)	(0.273)	(0.175)	(0.199)	(0.285)	(0.266)	(0.225)	(0.253)
female	-0.092	0.025	-0.068	-0.07	-0.035	-0.068	-0.078	0.171
	(0.07)	(0.108)	(0.096)	(0.125)	(0.107)	(0.112)	(0.081)	(0.107)
time x female	0.039	-0.158	-0.061	-0.114	0.033	-0.062	-0.113	-0.258
	(0.123)	(0.174)	(0.126)	(0.174)	(0.132)	(0.196)	(0.155)	(0.184)
treatment x female	0.003	-0.126	0.046	-0.318*	-0.151	0.007	-0.086	-0.266**
	(0.132)	(0.134)	(0.133)	(0.161)	(0.147)	(0.159)	(0.12)	(0.12)
time x treatment x	0.115	0.64**	0.205	0.518**	0.454**	0.303	0.322	0.668***
female	(0.171)	(0.255)	(0.164)	(0.221)	(0.214)	(0.279)	(0.192)	(0.217)
constant	0.242***	0.872***	0.162***	0.603***	0.685***	0.262***	0.788***	0.979***
	(0.175)	(0.154)	(0.16)	(0.138)	(0.144)	(0.127)	(0.145)	(0.148)
other control variables	yes	yes	yes	yes	yes	yes	yes	yes
impact on females	0.742***	0.944***	0.175	0.639***	0.698***	0.788***	0.385**	0.788***
	(0.184)	(0.228)	(0.171)	(0.189)	(0.236)	(0.238)	(0.172)	(0.244)
average impact	0.678***	0.612***	0.069	0.358**	0.456*	0.63***	0.214	0.445*
	(0.163)	(0.219)	(0.153)	(0.158)	(0.239)	(0.21)	(0.177)	(0.225)

Table presents estimates of nine Tobit regression models. Standard errors denoted in parentheses. Statistical significance at the 1, 5, and 10 percent levels denoted by ***, ** and *, respectively. Impact on females is the estimated sum of time x treatment + time x treatment x female. Number of observations is 893 students. Average impact (for both genders) is estimated using a separate Tobit model including time, treatment and time, treatment (average impact) and other control variables as regressors.

An average impact is also estimated and presented in Table 7. This average impact is estimated using a Tobit model but excluding the gender variables. Overall, the program had a strong positive impact in the three reading domains targeted by the intervention and four of the six other reading domains.

C. Internal validity and robustness checks

The effect size was estimated using five other methods to test the robustness of the model's estimates given the imbalance and school attrition described above; these results are not presented in this paper but available from the author on request. In the first method, the effect sizes were estimated using no control variables. In the second, the end-line data alone are used to estimate impact as this provides a larger sample size.

One reason for poor balance in the baseline may be the relatively small population of schools to draw on. Recent studies in the medical research field have dealt with this source of poor balancing (van Marwijk et al. 2008; Xu and Kalbfleisch 2010; Ravaud et al. 2009; Roux et al. 2011; Taft et al. 2011; Schwartz et al 2015; Leyrat et al. 2016). Leyrat et al. (2013) use a Monte Carlo simulation to assess the accuracy of several different methods including the use of covariates, weighting observations by the inverse of the probability of being selected into their respective treatment or control group (e.g.: Seaman and White 2013) and a direct adjustment by including this probability as a covariate. These latter two methods are the third and fourth methods used in this paper to test for robustness. Finally, Lee bounds (Lee 2009) are estimated using data aggregated at the school level to test whether school attrition may affect the results.

For all five methods, the results are similar to the estimates of the model. Only the estimates from the first method, the Tobit model without covariates, and the fifth method, school-level Lee bounds (without covariates) yielded notably smaller effect sizes. The remaining methods produced effect sizes similar to those presented in Table 7.

Two other issues may affect the internal validity of the estimates of impact. First, some contamination of control schools was reported, as Catholic schools in Madang Province received some special training in phonics. Second, because the data are repeated cross-sections of different cohorts, there are no data on student dropout or non-response. If the intervention affects whether students are present for the end-line data collection, then effect sizes may be biased.

D. External validity

Schools were randomly selected from a pre-defined population of schools that excluded very small schools, schools in highly remote areas, and those in dangerous areas. Attrition of schools from the Madang Province sample was a result of financial, logistical and weather issues. While these issues were unrelated to the treatment, if they reflect underlying characteristics of the schools that, in turn, affect the impact of the treatment, then this would introduce some bias. More generally, the effect sizes estimated in this paper may not be replicable in the more remote areas of the country or in those prone to the issues that led to the attrition of the schools in Madang.

5. Cost effectiveness

Benchmarking the impact of the Reader Booster Programme to other interventions helps assess the efficiency of the intervention and benefits of scaling up the intervention versus other types of interventions. The Abdul Latif Poverty Action Lab at the Massachusetts Institute of Technology compiles data on costs and impacts of several randomized impact evaluations in education. Table 8 presents the cost effectiveness of these programs measured as the impact on test scores in standard deviations per 100 USD cost. In their data, 2.278 standard deviations is the median impact per 100 dollars. The figures are not perfectly comparable. Tests differ in grade level and psychometric properties, but it provides a general range of cost effectiveness data with which to benchmark the Reading Booster Programme.

Table 8. Estimated impact of selected randomized interventions (standard deviations per 100 US dollars)

Intervention	SD/$100
Providing earnings information, Madagascar	118.338
Streaming by achievement, Kenya	34.784
Linking school cmte to local govt, Indonesia	34.624
Electing school cmte & linking to local govt, Indonesia	13.337
Teacher incentives (year 2), Kenya	6.291
Textbooks for top quintile, Kenya	3.563
Remedial education, India	3.069
Camera monitoring, India	2.278
Village-based schools, Afghanistan	2.126
Extra contract teacher + streaming, Kenya	1.971
Individually-paced computer assisted learning, India	1.551
Girls Scholarships, Kenya	1.384
Read-a-thon, Philippines	1.176
Minimum conditional cash transfers, Malawi	0.06
Contract teachers, Kenya	-0.299

Source: Abdul Latif Poverty Action Lab (2014)

The total cost for each year of the Reading Booster Programme was 794,243 PGK (250,549 USD) based on data from the World Bank project which supported the program. Because this program was implemented during regular teaching hours, there is no additional cost of teachers; these costs reflect training and distribution of materials as well as monitoring and evaluation. The program benefitted 4,272 students, yielding a cost of 186 PGK (59 USD) per student (World Bank 2016c). Table 9 presents the average impact of the program in standard deviations per 100 USD, which is calculated by dividing the impacts presented in Table 7 by 0.59. Per 100 USD, the impact of this program on the two targeted reading domains ranges from 1.04 to 1.16 standard deviations. For the other reading domains where a statistically significant effect was found, effect sizes range from 0.61 to 1.07 standard deviations per 100 USD. The reader booster program is most cost effective for girls; cost effectiveness, like effect size presented in Table 7, is higher in the five domains that have statistically higher effect sizes than males.

Table 9. Estimated impact of intervention on reading domains in standard deviations per 100 USD cost per student

	Domains targeted by the intervention		Other reading domains					
	Initial sounds	Familiar words	Letter names	Unfamiliar words	Reading comp.	Oral comp.	Dictation	Oral read. fluency
Effect per 100 USD	1.069***	0.519	-0.052	0.206	0.417	0.826*	0.107	0.204
males	(0.317)	(0.465)	(0.298)	(0.339)	(0.485)	(0.453)	(0.384)	(0.432)
Effect per 100 USD	1.264***	1.61***	0.298	1.09***	1.19***	1.343***	0.656**	1.343***
females	(0.314)	(0.388)	(0.291)	(0.322)	(0.403)	(0.405)	(0.294)	(0.416)
Average effect per 100	1.156***	1.044***	0.117	0.611**	0.778*	1.074***	0.365	0.759*
USD	(0.278)	(0.373)	(0.261)	(0.27)	(0.407)	(0.359)	(0.301)	(0.384)

Table presents average effects of the intervention in standard deviations per 100 USD; this calculated using the average effects of Table 7 and dividing by the cost per student in 100s USD (58.65 USD). Statistical significance at the 1, 5, and 10 percent levels denoted by ***, ** and *, respectively.

Compared to the data compiled by the Poverty Action Lab, this intervention's cost effectiveness lies towards the bottom end of the distribution. However, it is not clear how this intervention will affect test scores later in the students' schooling. The impact of the program may be amplified over time, as early reading skills are crucial to a child's literacy and future learning.

6. Conclusions

These findings provide evidence that a teacher training approach providing highly scripted lesson plans can improve basic reading skills in a low-cost, remedial course format, especially for girls. The Papua New Guinea curriculum provides teachers with flexibility over how they use their instructional time for language; this flexibility permitted the piloting and evaluation of the program. The Reader Booster Programme diverges from the curriculum's approach by providing teachers with a very specific teaching method and scripted lesson plans that they apply within the time allocated to language instruction.

A natural question is how much flexibility should teachers have within the curriculum in a developing country context? Highly structured approaches are appealing in contexts where teacher qualification and ability vary considerably. While in developed countries, the use of highly scripted lesson plans has received mixed reception, evaluations of interventions providing teachers with specific teaching methods and lesson plans have been shown to be successful in developing countries to improve early reading skills. The Reader Booster Programme adds to this evidence-base demonstrating an intervention that is formatted as a remedial course aimed at improving specific reading skills. The remedial course format is advantageous because it can be implemented without changes to the national curriculum and can be targeted to schools most in need. Its low cost is also important given the education budget constraints faced by Papua New Guinea and other developing countries.

One limitation of the Reader Booster Programme is the weaker effect on boys. It is not clear from the data collected in this study why this may be the case. Qualitative work would be beneficial to better understand this outcome; however, the program has already completed. If this approach is replicated in other countries, gender differences in the effects should be studied closely.

References

Abadzi, H. 2006. *Efficient learning for the poor: Insights from the Frontier of Cognitive Neuroscience.* Washington, DC: The World Bank.

Alexander, K.L., Entwisle, D.R., Horsey, C.S., 1997. From first grade forward: early foundations of high school dropout. *Sociology of Education* 70 (2), 87–107

August, D., and Shanahan, T. 2006. *Developing Literacy in Second-Language Learners: A Report of the National Literacy Panel on Language, Minority Children, and Youth*. Mahwah NJ USA: Lawrence Erlbaum Associates

Azariadis, C. and A. Drazen 1990. Threshold externalities in economic development. *The Quarterly Journal of Economics*. 105 (2): 501-526.

Basu, K. and J. Foster 1998. On measuring literacy. *Policy Research Working Paper Series*. No. 1997. Washington, D.C.: The World Bank

Butler, S.R., H. W. Marsh, M. J. Sheppard, and J. L. Sheppard 1985. Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology*, 77, 349-361

Chiappe, P., L. Siegel, and L. Wade-Woolley. 2002. Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading* 6(4): 369–400

Entwisle, D. R., K. L. Alexander, and L. S. Olson 2005. First grade and educational attainment by age 22 A New Story. *American Journal of Sociology*, 110 1458-1502

Fuchs, L., D. Fuchs, M.K. Hosp, and J. Jenkins. 2001. Oral Reading Fluency as an Indicator of Reading Competence: A Theoretical, Empirical, and Historical Analysis. *Scientific Studies of Reading* 5(3), 239–256.

Gove, A. and P. Cvelich. 2011. *Early Reading: Igniting Education for All. A report by the Early*

Grade Learning Community of Practice. Revised Edition. Research Triangle Park, NC: Research Triangle Institute.

Government of Papua New Guinea 2015. Pilot reading booster program in Western Highlands Province, Papua New Guinea. Unpublished manuscript.

Government of Papua New Guinea 2016. Madang Reading Booster Evaluation Report. Unpublished manuscript.

Hirano K, and Imbens G.W. 2001. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology Dec 2001*; 2(3):259–278.

Jimerson, S., Egeland, B., Sroufe, L.A., Carlson, B., 2000. A prospective longitudinal study of high school dropouts: examining multiple predictors across development. *Journal of School Psychology* 38 (1), 525–549

Leyrat, C., A. Caille, A. Donner and B. Giraudeau 2013. Propensity Scores Used for Analysis of Cluster Randomized Trials with Selection Bias: a Simulation Study. *Statistics in Medicine · August 2013*.

Lee, D. S. (2009). Training, wages, and sample selection: estimating sharp bounds on treatment effects. *The Review of Economic Studies*. 76(3): 1071-1102

Leyrat, C., A. Caille, Y. Foucher and B. Giraudeau 2016. Propensity score to detect baseline imbalance in cluster randomized trials: the role of the c-statistic. *BMC Medical Research Methodology (2016)* 16:9DOI 10.1186/s12874-015-0100-4

Linan-Thompson, S., and S. Vaughn. 2007. *Research based methods of reading instruction for English language learners: Grades K–4*. Alexandria, VA: Association for Supervision and Curriculum Development

National Institute for Child Health and Human Development 2000. *Report of the National Reading Panel. Teaching Children to Read: An Evidence-based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction*. (NIH Publication No. 00-4754). Washington, DC: National Institutes of Health

National Reading Panel. 2000. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Washington, DC: National Institute of Child Health and Human Development.

OECD 2014. *PISA 2012 Results: What Students Know and Can do: Student Performance in Mathematics, Reading and Science (Volume I) [Revised edition February 2014]*. Paris: OECD

PASEC 2015. *PASEC 2014. Education System Performance in Francophone Sub-Saharan Africa. Competencies and learning factors in primary education.* Dakar: CONFEMEN

Piper, B. and M. Korda 2010. *EGRA Plus: Liberia. Program Evaluation Report.* Research Triangle Park, N.C. U.S.A: RTI International

Piper, B. 2008. Integrated Education Program Impact Study of SMRS Using Early Grade Reading Assessment in Three Provinces in South Africa. Research Triangle Park, N.C. U.S.A: RTI International

Pressley, M., 1998. *Reading Instruction That Works: The Case for Balanced Teaching*. New York: The Guilford Press

Ravaud P, Flipo R, Boutron I, Roy C, Mahmoudi A, Giraudeau B, Pham T. ARTIST (osteoarthritis intervention standardized) study of standardised consultation versus usual care for patients with osteoarthritis of the knee in primary care in france: pragmatic randomised controlled trial. *BMJ (Clinical research ed.)* 2009; 338:b421.

Rosenzweig, M. R. 1995. Why are there returns to schooling? *The American Economic Review*. Vol. 85, No. 2: 153-158

Roux C, Giraudeau B, Rouanet S, Dubourg G, Perrodeau E, Ravaud P. Monitoring of bone turnover markers does not improve persistence with ibandronate treatment. Joint, Bone, Spine: *Revue Du Rhumatisme* Jun 2011; doi:10.1016/j.jbspin.2011.05.001.

Scarborough, H. S. 2002. Connecting Early Language and Literacy to Later Reading (Dis)abilities: Evidence, Theory, and Practice. In: Dickinson, D.K. and S.B. Neuman. *Handbook of Early Literacy Research (vol. 1)*. Edited by. New York: The Guilford Press: 97-110

Schwartz R, Vigo Á, Dias de Oliveira L, Justo Giugliani ER (2015) The Effect of a Pro-Breastfeeding and Healthy Complementary Feeding Intervention Targeting Adolescent Mothers and Grandmothers on Growth and Prevalence of Overweight of Preschool Children. *PLoS ONE* 10(7):e0131884. doi:10.1371/journal.pone.0131884

Seaman, S. R. and I. R. White (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3): 278-295.

Spratt, J., S. King, and J. Bulat 2013. *Independent Evaluation of the Effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) Program in Mali. End-line Report.* Research Triangle Park, N.C. U.S.A: RTI International

Sprenger-Charolles, L. 2004. Linguistic Processes in Reading and Spelling: The Case of Alphabetic Writing Systems: English, French, German and Spanish. In: Nunes, T. and P. Bryant (Eds.) *Handbook of Children's Literacy*. Dordrecht, the Netherlands: Kluwer Academic Publishers: 43–66

Snow, C.E., Burns, M.S., and Griffin, P. 1998. *Preventing Reading Difficulties in Young Children*. National Academy Press, Washington, DC

Taft AJ, Small R, Hegarty KL, Watson LF, Gold L, Lumley JA. Mothers' Advocates in the community (MOSAIC)–non-professional mentor support to reduce intimate partner violence and depression in mothers: a cluster randomised trial in primary care. *BMC public health* 2011; 11:178, doi:10.1186/1471-2458-11-178.

van Marwijk HW, Ader H, de Haan M, Beekman A. Primary care management of major depression in patients aged >= 55 years:. *The British Journal of General Practice* Oct 2008; 58(555):680–687, doi:10.3399/bjgp08X342165.

Wolf, M. 2007. *Proust and the Squid: The Story and Science of the Reading Brain*. New York: Harper Collins

World Bank 2011. *Project appraisal document on a proposed catalytic fund grant in the amount of US$19.2 million to Papua New Guinea for a reading education (READ PNG) project.* Washington, D.C..: The World Bank

World Bank 2012a. *How well are Tongan children learning to read?* Washington, D.C.: The World Bank

World Bank 2012b. *How well are Ni-Vanuatu children learning to read in English?* Washington, D.C.: The World Bank

World Bank 2012c. *How well are Ni-Vanuatu children learning to read in French?* Washington, D.C.: The World Bank

World Bank 2014a. *East New Britain (ENB) Early Grade Reading Assessment (EGRA) Survey. 2012 Diagnostic Results Report*. Washington, D.C.: The World Bank

World Bank 2014b. *Madang Early Grade Reading Assessment (EGRA) Survey. 2011 Diagnostic Results Report*. Washington, D.C.: The World Bank

World Bank 2014c. *National Capital District (NCD) Early Grade Reading Assessment (EGRA) Survey. 2012 Diagnostic Results Report*. Washington, D.C.: The World Bank

World Bank 2014d. *Western Highlands Province Early Grade Reading Assessment (EGRA) Survey. 2013 Diagnostic Results Report*. Washington, D.C.: The World Bank

World Bank 2016a. *World DataBank – World Development Indicators*. Accessed December 2016, http://databank.worldbank.org/data/reports.aspx?source=2&country=PNG

World Bank 2016b. *Data Query*. Accessed December 2016. http://datatopics.worldbank.org/education/wDataQuery/QFull.aspx

World Bank 2016c. *Implementation completion and results report on a grand in the amount of US$19.2 million to the Independent State of Papua New Guinea for a reading education (READ-PNG) project*. Washington, D.C.: The World Bank

World Bank (2017a). *Tuvalu Early Grade Reading Assessment (TuEGRA): results report*. Washington, D.C.: The World Bank

World Bank (2017b). *Kiribati Early Grade Reading Assessment (KiEGRA): results report*. Washington, D.C.: The World Bank

Xu, Z. and J.D. Kalbfleisch 2010. Propensity score matching in randomized clinical trials. *Biometrics*. 2010 Sep; 66(3):813-23. doi: 10.1111/j.1541-0420.2009.01364.x