NORC WORKING PAPER SERIES

Improving Teaching Quality through Training: Evidence from the Caucasus

WP-2017-006 | MAY 2017

PRESENTED BY:
NORC at the University of Chicago
55 East Monroe Street
30th Floor
Chicago, IL 60603
Ph. (312) 759-4000

AUTHORS:
Alejandro Ome, NORC at the University of Chicago
Alicia Menendez, University of Chicago and NORC at the University of Chicago
Elise Le, World Bank

AUTHOR INFORMATION

Alejandro Ome
NORC at the University of Chicago
55 East Monroe suite 3100
Chicago, IL 60603
(312) 357-3884
ome-alejandro@norc.org

Alicia Menendez
University of Chicago and NORC at the University of Chicago
ali.menendez@gmail.com

Elise Le
World Bank
hle2103@gmail.com

Table of Contents

Table of Tables

Table of Figures

Abstract

This study analyzes the effect of the Georgian Primary Education Project, an initiative that provided teacher training to 122 schools in Georgia between 2013 and 2015. We use Value-Added models to estimate the effect of the program on math and Georgian test scores of students who were in grades 1-4 in 2013. We find positive effects for both math and Georgian for students who were in grades 1 to 3 at baseline, but no effects for those in grade 4 at baseline. We do not find any effects on Georgian as a second language among students in ethnic minority schools.

1. Introduction

Teaching quality is arguably the most important school-based factor determining student achievement. Teaching quality can be modified via three channels: improving the type of teachers working in schools; providing incentives to exert greater teacher effort (monetary or non-monetary); and improving the quality of teaching through training and professional development. Between 2013 and 2015, a program conducted in the Republic of Georgia followed the third channel – teacher training – to improve teaching quality. In this study, we use Value-Added models to analyze the effect of this program on student achievement, measured by math and reading test scores.

There is mixed empirical evidence on whether teacher training improves student achievement. Glewwe et al. (2013) review high-quality studies conducted between 1990 and 2010 on teacher training (and other types of school interventions) and find mixed results for this type of programs. As Bruns & Luque (2015) point out, given the heterogeneity of teachers' characteristics and training programs' content and intensity, the lack of consensus on the effects of teacher training is hardly surprising.

However, some interventions that follow a comprehensive approach to teacher training show positive impacts. For example, Chay et al. (2005) use a regression discontinuity design to evaluate the Chilean P-900, an intervention targeting low-performing schools that provided teacher training, infrastructure improvement, textbooks and other instruction materials, and tutoring for low-performing students. The evaluation documents positive effects on students' test scores of 0.2 standard deviations. More recently, Piper & Korda (2011) evaluate a program in Liberia that provided teacher training through a combination of capacity-building workshops, on-going regular feedback, as well as other community outreach activities. Using randomization as their identification strategy, the authors find that the program improves reading scores by 0.79 standard deviations. Menendez and Dayaratna (2016) evaluate a similar intervention in Uganda using an experimental design but find small effects on reading fluency among third graders exposed to the program since the beginning of their primary education. Lucas et al. (2014) compare the effect of fielding a teacher-training intervention in Uganda and Kenya. They use experimental designs in each country to evaluate a program that included teacher training, instruction materials, and ongoing mentoring for teachers. Lucas and her coauthors find significant effects for Uganda of approximately 0.2 standard deviation, but much smaller effects for Kenya. Oliveira & Carnoy (2015) use a triple-difference approach to evaluate *Pacto pela Alfabetizacao na Idade Certa*, an early grade reading program in Brazil that provided teacher training and reading materials to schools, combined with monetary incentives based on student performance in standardized exams; they find effects of 0.08 and 0.14 standard deviations for Portuguese and math, respectively.

What these programs have in common is a comprehensive approach to teacher professional development. These interventions do not simply provide teacher training but also offer a series of teacher-support resources, including regular feedback and teaching materials.

The Georgia Primary Education Project (G-PriEd) took a similar approach to teacher-professional development. In addition to teacher training, the program provided ongoing support for teachers and principals. It supplied instructional materials including leveled supplementary readers, students' newspapers, and math manipulatives.[1] In addition, to help teachers check their own teaching quality and inform them on their students' performance, the program equipped teachers with students' formative assessments tools. Finally, to foster accountability and transparency as an external check on teaching quality, the program created school report cards for principals with information from school performance on training participation, teacher tests, use of project methodology in the classroom, and other project activities.

The main challenge in evaluating G-PriEd is that schools self-selected into the program. Specifically, the Georgian Ministry of Education and Science (MES) invited schools to apply to the program through a promotional campaign. Of a total of 817 applications received, 122 pilot schools were then chosen on a first-come first-served basis to participate. As a comparison group, 119 schools were randomly chosen from the pool of schools that did not apply. Because the schools that self-selected into applying might be different from those that did not, it is difficult to isolate the effects of G-PriEd from these other

[1] A reader is a book that provides supplemental reading materials; a manipulative is an object design to teach mathematical concepts.

differences. For example, the schools that applied might be the ones that were most interested in improving test scores to begin with. In such a scenario, any difference in the teaching quality of the treatment group may be due to their own efforts instead of G-PriEd.

To tackle this selection problem, we exploit the fact that data on the same students was collected at baseline and endline, and estimate a Value-Added Model (VAM) to evaluate the impact of the program. The key feature of VAM is the inclusion of a lagged achievement measure (baseline) as a control variable. For VAM to identify the causal impact of the program, the underlying assumption is that baseline test scores are sufficient to characterize the cognitive ability of students at that moment (for a discussion on this type of models see Todd & Wolpin, 2003). While we cannot test whether this assumption holds in the context of G-PriEd, a growing literature shows that VAM can replicate experimental parameters of schooling interventions, specifically in the context of charter schools in the U.S. (Abdulkadiroglu et al., 2011; Deutsch, 2012; and Deming, 2014).

We find that G-PriEd has significant positive effects on students' achievement. For math, we estimate an average impact of 0.27 standard deviations and for reading Georgian as a native language, an effect of 0.15 standard deviations. For Georgian as a Second Language (GSL) no significant effects are found. The lack of results for GSL students may be due to the fact that these teachers received less training than they were initially scheduled for due to budgetary restrictions. We also explore treatment heterogeneity across gender and baseline test scores. We do not find strong evidence that the program had differential effects across these dimensions.

This study contributes to the growing literature on comprehensive teacher training as a strategy to improve students' achievement in developing countries. To the best of our knowledge, this is the first rigorous evaluation of a teacher-training program in Eastern Europe, so the results from this study may be especially relevant for countries in the region that are trying to improve school quality.

This paper has seven sections including this introduction. In the next section, we describe the G-PriEd program. Section 3 presents the data and our methodological approach. Section 4 describes the main results and section 5 the heterogeneity analyses. In section 6 we discuss the results. Section 7 contains the conclusion.

2. The G-PriEd program

Georgia is a country located east of the Black Sea, with a population of 3.7 million and a per capita GDP of US 9,163 in 2014 (PPP). Despite of being a middle-income country, Georgia struggles to improve the quality of the education provided and learners' outcomes. For example, the results of the Program for International Student Assessment (PISA) 2009 indicate that less than 40 percent of 15-year-old students reach reading proficiency level. Georgia compares very poorly to most participant nations, not only on reading but also on mathematics and science (Walker, 2011).

G-PriEd was a pilot project funded by USAID and implemented by Chemonics International in collaboration with the MES, which aimed to improve primary students' skills in reading and mathematics. The program took a comprehensive approach to provide multiple services to teachers and schools. First, principals and teachers were trained in instructional practices in reading and math. Furthermore, teachers received continuous support through school-based Teacher Learning Circles (TLC). During these sessions, teachers discussed student progress, test scores, and brainstormed solutions to any challenges. Also, teachers received support through classroom visits from national trainers who gave teachers constructive feedback as a result of the observations. Second, G-PriEd provided teachers with a student-assessment tool that equipped teachers with real-time information that they could use to adapt teaching practices. Third, schools were also provided with supplementary leveled readers for each grade as well as other instructional materials. Finally, the project aimed to foster accountability and transparency by creating school report cards.

All 122 pilot schools received educational equipment and math manipulatives in spring 2013, and in October 2013 and March 2014 all 122 pilot schools received the supplementary leveled readers.

In spring 2013, G-PriEd trained reading and math teachers from grades 1-6 of treatment schools. In fall 2013 and spring 2014, due to budget restrictions, the G-PriEd team only trained reading and math teachers from grades 1 to 4 and in Georgian schools only, resulting in 103 schools trained out of a total of 122. However, G-PriEd continued to train principals as well as the TLC facilitators from all treatment schools. In fall 2014 and spring 2015, the training resumed with trainings of teachers from all primary grades, 1 to 6, in both Georgian and ethnic minority schools. Table 1 shows the total number of days of training offered by G-PriEd by grade and subject in Georgian and ethnic minorities schools. One day of training consisted of a 6-hour session..

3. Data and Empirical Methods

Among all applications to the program, 122 pilot schools were chosen on a first-come first-served basis. More specifically, all public schools in the country were divided in 43 blocks, according to their size[2] (small, 1-299 students; medium, 300-599 students; and large schools, over 600 students), language of instruction (Georgian and Non-Georgian) and 12 geographic clusters (11 administrative regions of Georgia, including its breakaway region of Abkhazeti). In proportion to the number of students in each block, the project team established that between 1 and 15 schools per block would be selected for treatment. Once the number of schools per block was defined, schools were selected on a first-come first-served basis. The project team selected the same number of schools per block (when possible) as comparison groups.

To evaluate G-PriEd, we collected data in spring 2013, before the program started, with samples of students from 122 treatment and 119 comparison schools from grades 1 to 6 using G-PriEd's own math and Georgian diagnostic assessment tools. Math exams were translated into Russian, Azeri and Armenian to assess students in schools with these languages of instruction. For Georgian, two types of exams were fielded, one for Georgian native speakers and another for Georgian as a Second Language (GSL), for the schools where the language of instruction was Armenian, Azeri or Russian.

In each school, a sample of 1 to 6 students per grade was randomly selected to take the test, depending on school size. With this sampling strategy, the baseline target sample size consisted of 1,710 students from pilot schools and 1,557 students from the control schools for a total of 3,267.

In spring 2015, students from grades 1-6 from the same schools visited at baseline were surveyed. With the objective of constructing a panel of students, we aimed to survey the same students that were surveyed at baseline. Given that two years had elapsed since the baseline in spring 2013, only those students who were attending grades 1 to 4 at baseline were attending the grades we surveyed in 2015 (repetition rate is negligible in Georgia and in fact, we do not have any case in our sample). Of all the students surveyed in 2013, a total of 2,179 were in grades 1-4. In 2015 we tried to interview these students in grades 3-6 and surveyed approximately 80 percent of them. We could not survey the remaining 20 percent because they

[2] For this stratification the school size included all students in the school, that is, from grades 1 to 12.

were either attending a different school at endline or absent the day of data collection. As a result, the panel of students contains baseline and endline data for 1,696 students.

Table 2 shows baseline descriptive statistics for schools. The data for schools come from the Georgian Education Management Information Systems (EMIS), and the available variables are: number of teachers and students, number of certified teachers, and number of classrooms. Panel A in Table 2 shows results for schools where Georgian is the language of instruction, or “regular” schools. Teacher/student ratio is 0.27 for treatment schools and 0.23 for comparison schools, but the difference is not significant. Class size is also slightly higher for treatment than for comparison schools. The percentage of certified teachers is higher for treatment than for comparison schools, and in this case, the difference is statistically significant. Panel B in Table 2 shows results for ethnic minority schools. For these schools the sample is much smaller than for regular schools, and no difference is statistically significant. It is worth noting, however, that compared to regular schools, ethnic minority schools have a much lower percentage of certified teachers. In fact, while in regular schools 20 percent of teachers are certified, on average, in ethnic minority schools only 4 percent are. This suggests that regular schools are better staffed than ethnic minority schools.

Table 3 shows summary statistics for all the students in the panel sample. In this case, the available data are limited to the sex and age of the surveyed students. There are no major differences between treatment and comparison schools for these two characteristics. The last three rows of the table show the results for test scores. Note that test scores have been transformed such that the mean is 0 and the standard deviation is 1. Students in treatment schools outperform their comparison counterparts in math and reading, although only the difference for reading is statistically significant. For GSL, on the other hand, students in comparison schools have higher test scores on average than students in treatment schools, however the difference is not significant.

Overall, schools in treatment and comparable groups seem relatively different at baseline. The fact that we find significant differences between treatment and control schools in the fraction of certified teachers (in regular schools) suggests that schools in the treatment group may be better staffed than schools in the comparison group. Similarly, the documented significant differences for reading test scores indicate that students in treatment schools have higher achievement than students in comparison schools.

These differences suggest that comparing endline test scores between treatment and comparison students would be a biased estimate of the treatment effect. Such an approach will confound the causal impact of the program with the differences already observed at baseline. To tackle this identification problem, we make use of our panel of students and apply a Value-Added model (VAM).[3]

The key feature of VAM is the inclusion of a lagged achievement measure (baseline) as a control variable. The identification assumption underlying the VAM is that baseline test scores are a sufficient statistic to characterize the cognitive ability of students at baseline. Mathematically, for each student *i* we estimate:

$$A_{i1} = \alpha + \beta D_i + \gamma A_{i0} + \mathbf{x}_\mathbf{i}'\boldsymbol{\delta} + u_{i1} \qquad (1)$$

where, A_{i1} and A_{i0} are test scores for student *i* at endline (1) and baseline (0), respectively; D_i is a dummy variable for treatment status; $\mathbf{x}_\mathbf{i}$ is a vector of characteristics at the student level, specifically, age at baseline and gender, as well as characteristics at higher levels of analysis, in particular, categorized student/teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for 11 geographic regions; u_{i1} is an error term and α, β, γ and $\boldsymbol{\delta}$ are parameters to be estimated. The parameter of interest is β, which captures the effect of the program on endline test scores. To correct for within-school error correlation, standard errors are clustered at the school level.

We use this same framework to explore treatment heterogeneity across student gender and baseline tests scores. Specifically, to analyze the effect of the program by gender, we split the sample and estimate separate effects for boys and girls. To analyze treatment heterogeneity across baseline test scores, we add a third-order polynomial on baseline test scores interacted with the treatment dummy to Equation (1), and plot the treatment effect for the entire baseline test scores distribution; this way, we can analyze what the estimated treatment effect is depending on students' baseline test scores.

4. Results

Table 4 shows VAM results for math. For 3rd graders at endline, the effect of G-PriEd is 0.41 standard deviations. For 4th graders it is 30 percent and for 5th graders it is 30 percent. These effects are statistically significant and relatively large in the context of schooling interventions. For 6th graders the parameter is positive but small and not significant. Overall, we find that G-PriEd significantly improves students' math

[3] Another alternative to address this identification issue would have been to use schools that applied to the program but not soon enough to be selected as the comparison group instead of drawing a sample from the pool of schools that did not apply at all. However, in the early stages of the program, it was decided that comparison schools would be drawn from the schools that did not apply to the program.

test scores. If we run a regression pooling students in all four grades, the average effect is 0.27 standard deviations. Perhaps not surprisingly, we can also see that, with no exception, the baseline test scores of both math and reading are positively correlated with the endline math test score.

Table 5 presents the results for reading Georgian as a native language. For students in 3rd grade at endline there is a positive and significant effect of 0.27 standard deviations. For students in 4th grade the effect is also positive but not significant at standard levels of confidence (p-value=0.09). For students in grade 5 at endline there is a positive and significant effect of 0.2 standard deviations. The effect for grade 6 at endline is negative but very small and not significant. The pooled estimate indicates a significant effect of 0.15 standard deviations.

Finally, Table 6 presents the results for reading Georgian in ethnic minorities' schools, or Georgian as a Second Language (GSL). No coefficient is statistically significant and only one is positive.

As it was already mentioned, at endline it was not possible to survey a fifth of the baseline sample. To try to determine whether attrition is biasing the presented results, as a sensitivity exercise we modeled the attrition process using baseline characteristics, and ran all regressions using inverse probability weights to control for attrition.[4] The results hardly change with respect to the unweighted regressions just discussed.

5. Treatment Heterogeneity

In this section we analyze treatment heterogeneity across two dimensions: gender and baseline test scores. Given the small sample available for GSL, in this section we focus on results for math and reading Georgian as a native language only.

Table 7 shows results for boys and girls by subject. There are no apparent differences by students' sex. For math, the treatment effect for both boys and girls is significant, and while the point estimate for girls is higher than for boys, the difference is relatively small (roughly equal to the standard errors). For reading, the coefficient is only significant for girls, but the difference between the two coefficients is only

[4] Specifically, the baseline variables included in the attrition model were the treatment dummy, students' gender, a categorical variable for school size, dummy variables for the three minorities (Russian, Azeri and Armenian), student/teacher ratio, class size, percent of certified teachers, math baseline score (because reading scores are not comparable between Georgian and GSL, only math scores were used to model attrition) and dummies for grades 1-4 at baseline.

about as large as the standard errors. This shows that G-PriEd did not have strong differential effects by students' sex, albeit the point estimates are slightly larger for girls than for boys.

To test treatment heterogeneity by baseline achievement level, we estimate Equation (1) adding a third-order polynomial on baseline test scores interacted with the treatment dummy. Figure 1 shows treatment effects for math by baseline test scores using this model, as well as the 95 percent confidence interval. At the tails of the baseline test scores distribution, the confidence intervals blow up so we trimmed the x-axis to a (-2, 2) range, where roughly 95 percent of the data fall. At the lower tail of the baseline test score distribution (roughly between -2 and -1), the estimate treatment effects are not significant, but the effects grow with the baseline test score, and at the right tail (between 1 and 2) the effect is positive and relatively large, approximately 0.4 standard deviations. This indicates that G-PriEd had a greater effect on students who were above the average at baseline, suggesting that the program did not help closing achievement gaps observed at baseline.

Results for reading Georgian as a native language are displayed in Figure 2. A similar, but not as strong, pattern can be observed. In this case, the marginal effects are also larger at the higher tail of the baseline test scores than at the lower tail, but the differences are not significant. However, the estimated effects of G-PriEd are only significant for students scoring above average at baseline.

One of the teaching practices introduced by G-PriEd involved grouping students by their performance within the same classroom. It is possible that this mild form of tracking has led to the differential effects of G-PriEd across students' baseline abilities. Further research should address this type of distributional impacts of G-PriEd and other programs like it.

6. Discussion

It may be surprising that G-PriEd had a positive effect on math and reading (Georgian as a native language) achievement on students who were in grades 3 to 5 at endline, but no effect on students who were in grade 6 at endline. The lack of results for the 6th grade may be explained by two main factors. First, due to how primary grade teachers are assigned to classes in Georgia, 5th and 6th graders may have had less trained teachers. In general, teachers in grades 1 through 4 follow their cohort of students as they move up grades. In other words, a teacher starts with a cohort in grade 1, follows that cohort all the way to grade 4, and once they reach grade 4, the teacher returns to grade 1 to follow a new cohort of students. In grades 5 and 6, reading and math are taught by different teachers, and these teachers do not necessarily follow students over grades; this depends on the school principal and is beyond the control of the G-PriEd project. Given this system, students in grades 1 to 4 may have been taught by teachers with more accumulated training than their counterparts in 5th and 6th grades.

Second, as explained in Section 2, four G-PriEd training waves were deployed between 2013 and 2015, but 5th and 6th grade teachers were trained in only two of them (in 2013 and 2015), due to budget constraints. This may have reduced effectiveness of 5th and 6th grade teachers compared to teachers in grades 1-4, and ultimately reduced students' achievement gains in those grades.

While teachers in both grades 5 and 6 are affected by the two aforementioned issues (rotation and the lower dosage than teachers in grades 1-4), the students in the treatment group who were in 5th grade at endline were less exposed to these issues than the students that were in 6th at endline. This difference could explain why we still find effects for 5th graders but not for 6th graders.

The lack of results for GSL is possibly due to similar considerations. In effect, we have the same dosage shortage observed by 5th and 6th grade teachers of Georgian students. On the other hand, and perhaps more importantly, during program implementation, training of ethnic minority school teachers proved more challenging than that of Georgian school teachers. According to the project implementer, it was difficult to identify qualified staff to translate the training materials and supplementary reading materials, and some teachers from ethnic minority schools did not have a mastery of the Georgian language that was adequate to understand the trainings (and presumably teach Georgian as a second language).

7. Conclusions

In this study we present the estimated effects of the Georgia Primary Education Project on students' learning outcomes using a Value-Added model. We find positive effects in mathematics for students who were in grades 1 to 3 at the beginning of the program and were assessed two years later when they were in grades 3 to 5, respectively. We find no effects for students who were in grade 4 when the program started and in grade 6 at endline. In the case of reading, we find positive effects for students who were in grades 3 to 5 when assessed at endline (albeit for grade 4 at endline the effect is significant only at 10 percent of confidence). As was the case for math, we do not find effects for students attending grade 6 at endline. In addition, we do not find statistically significant effects on reading Georgian as a second language among students in ethnic minority schools.

As we discussed, 5th and 6th grade teachers and teachers of students who are not Georgian native speakers did not receive the full training G-PriEd intended to have, and this may have inhibited the benefits of the program. In addition, training teachers in minority schools proved to be more challenging than training Georgian school teachers. It was difficult to find staff qualified to translate the project materials, and some teachers from ethnic minority schools were not proficient in Georgian, which compromised their ability to understand G-PriEd trainings fully and, almost certainly, their ability to provide high-quality GSL lessons to their students. Furthermore, as it was shown, certified teachers were much less prevalent in ethnic minority schools than in Georgian schools, which suggests that teaching quality may be lower in ethnic minority schools than in regular Georgian schools.

These findings have a few important implications for the design of teacher-training programs similar to G-PriEd. First, even comprehensive programs like this may be found challenging to improve teaching quality if ‘baseline’ teaching quality is too low. The fact that no results were found for GSL seems to be linked to the quality of teachers in those schools, so perhaps programs with this level of intensity are not enough for correcting this type of problems.

Second, dosage matters. Students in 6th grade in 2015 were the most affected by the lack of teacher training for 5th and 6th grade teachers in the middle of the program, and it is for these students that we do not observe any results for math and Georgian as a native language. A similar statement can be made about students taking the GSL. This suggests that is very important to provide recurrent training in the context of this type of programs. It also raises the question of whether the observed gains can be preserved once the program ends. Future research should address this key question in the context of this program and other similar efforts.

That being said, in general our findings are encouraging. The estimates indicate that, for math and reading Georgian as a native language, students in grades 3 to 5 at endline benefited from the program. This corroborates the findings from other research that have documented that comprehensive teacher training can positively affect students’ achievement (Chay et al., 2005; Piper & Korda, 2011; Lucas et al., 2014; Oliveira & Carnoy, 2015). Many developing countries spend substantial resources on teacher training (Bruns & Luque, 2015), the evidence presented here and in the studies mentioned indicate that comprehensive teacher training that incorporates not only lessons for teachers but constructive feedback and pedagogical materials among other components, has the potential to affect students’ achievement in a positive manner.

References

Abdulkadiroglu, Atila, Joshua D. Angrist, Susan M. Dynarski, Thomas J. Kane, and Parag Pathak (2011). Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *Quarterly Journal of Economics* 126(2)699-748.

Bruns, Barbara and Javier Luque (2014). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, DC: World Bank Group.

Chay, Kenneth Y., Patrick J. McEwan, and Miguel Urquiola. (2005). The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools. *American Economic Review*, 95(4)1237–1258.

Deming, David (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review: Papers & Proceedings* 104(5)406–411.

Deutsch, Jonah (2012). Using School Lotteries to Evaluate the Value-Added Model. University of Chicago Working Paper.

Glewwe, Paul, Eric A. Hanushek, Sarah Humpage, and Renato Ravina (2013). School Resources and Educational Outcomes in Developing Countries: A Review of the Literature from 1990 to 2010. In *Education Policy in Developing Countries*, ed. Paul Glewwe, 13–64. Chicago: University of Chicago Press.

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware, and Moses Oketch (2014). Improving early-grade literacy in east Africa: experimental evidence from Kenya and Uganda. *Journal of Policy Analysis and Management* 33(4)950–976.

Menendez, Alicia and Varuni Dayaratna (2016). Uganda: Impact Evaluation of SHRP Reading Program, Year 3. NORC at the University of Chicago.

Muralidharan, Karthik and Venkatesh Sundararaman (2010). The impact of diagnostic feedback to teachers on student learning: experimental evidence from India. *The Economic Journal* 120 (August) F187–F203.

Oliveira, Leandro & Martin Carnoy (2015). The Effectiveness of an Early-Grade Literacy Intervention on the Cognitive Achievement of Brazilian Students. *Educational Evaluation and Policy Analysis*, 37(4)567-590.

Piper, B., & Korda, M. (2011). EGRA Plus: Liberia, program evaluation report. Research Triangle Park, NC: RTI International.

Todd, Petra E. and Kenneth I. Wolpin (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, 113 (February)3-33.

Walker, Maurice (2011). PISA 2009 Plus Results: Performance of 15-year-olds in reading, mathematics and science for 10 additional participants. Melbourne: ACER Press.

Tables

Table 1. Number of days of training offered per year, grade and subject

	Georgian Schools				Ethnic Minority Schools		
	Grades 1-4		Grades 5-6		Grades 1-6	Grades 1-4	Grades 5-6
	Math	Reading	Math	Reading	GSL	Math	Math
April-June 2013	3 days	4 days	3 days	3 days	4 days	3 days	3 days
Nov-Dec 2013	2 days	2 days	none	none	none	none	none
March-April 2014	2 days	2 days	none	none	none	none	none
Oct 2014-Feb 2015	4 days	3 days	7 days	5 days	5 days	7 days	7 days

Source: G-PriEd

Table 2. Baseline descriptive statistics - Schools

	Treatment	Comparison	Diff	(se)
A. Regular Georgian schools				
Student/teacher ratio	0.27	0.23	0.04	(0.03)
Class size (students)	12.8	12.7	0.1	(1.2)
Percent of certified teachers	21.8	17.8	4.0	(1.8)
N	103	101		
B. Ethnic minority schools				
Student/teacher ratio	0.24	0.22	0.0	(0.1)
Class size (students)	16.9	16.6	0.3	(4.8)
Percent of certified teachers	4.7	3.0	1.7	(2.2)
N	19	18		

Note: Sample sizes for class size are a little smaller due to item-specific missing data (N=194 for regular Georgian schools and N=34 for ethnic minority schools).

Source: EMIS data for 2013.

Table 3. Baseline descriptive statistics - Students

	Treatment	Comparison	Diff	(se)
Student characteristics (N=1696)				
Female (%)	47.5	48.8	-0.01	(0.02)
Age at baseline	8.0	8.0	0.0	(0.1)
Test scores				
Math N=(1696)	0.08	0.00	0.08	(0.05)
Reading (N=1472)	0.09	-0.04	0.13	(0.05)
GSL (N=224)	0.02	0.15	-0.13	(0.13)

Note: Only students that are observed both at baseline and endline are included.

Source: G-PriEd data for 2013.

Table 4. Effect of G-PriEd on math test scores

	Grade at endline				
	3rd	4th	5th	6th	Pooled
G-PriEd	0.41***	0.30***	0.30**	0.072	0.27***
	(0.10)	(0.089)	(0.10)	(0.097)	(0.059)
Baseline math	0.30***	0.47***	0.48***	0.52***	0.43***
	(0.081)	(0.052)	(0.080)	(0.074)	(0.033)
Baseline reading	0.44***	0.34***	0.37***	0.29***	0.36***
	(0.072)	(0.051)	(0.064)	(0.054)	(0.028)
Obs	412	430	423	431	1696

Note: All specifications include students' age, gender and special education status, categorized student-teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for language of test and 11 regions. Pooled regression includes cohort fixed-effects. Standard errors clustered at the school level in parentheses.

Source: G-PriEd and EMIS data for 2013 and 2015.

Table 5. Effect of G-PriEd on reading test scores (Georgian as native language)

	Grade at endline				
	3rd	4th	5th	6th	Pooled
G-PriEd	0.27*	0.15	0.20*	-0.014	0.15*
	(0.11)	(0.091)	(0.082)	(0.078)	(0.060)
Baseline math	0.18**	0.26***	0.25***	0.16***	0.21***
	(0.070)	(0.051)	(0.056)	(0.048)	(0.028)
Baseline reading	0.51***	0.39***	0.46***	0.48***	0.46***
	(0.059)	(0.053)	(0.051)	(0.042)	(0.030)
Obs	357	372	366	377	1472

Note: All specifications include students' age at baseline, gender and special education status, categorized student-teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for 11 regions. Standard errors clustered at the school level in parentheses. Pooled regression includes cohort fixed-effects.

Source: G-PriEd and EMIS data for 2013 and 2015.

Table 6. Effect of G-PriEd on reading test scores (Georgian as second language)

	Grade at endline				
	3rd	4th	5th	6th	Pooled
G-PriEd	-0.18	-0.11	0.034	-0.14	-0.14
	(0.21)	(0.19)	(0.21)	(0.20)	(0.11)
Baseline math	0.056	0.51***	0.11	0.25*	0.18**
	(0.26)	(0.079)	(0.11)	(0.11)	(0.053)
Baseline GSL	0.61**	0.68**	0.70**	0.31	0.60***
	(0.21)	(0.21)	(0.23)	(0.17)	(0.11)
Obs	55	58	57	54	224

Note: All specifications include students' age, gender and special education status, categorized student-teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for 11 regions. Standard errors clustered at the school level in parentheses. Pooled regression includes cohort fixed-effects.

Source: G-PriEd and EMIS data for 2013 and 2015.

Table 7. Effect of G-PriEd by gender

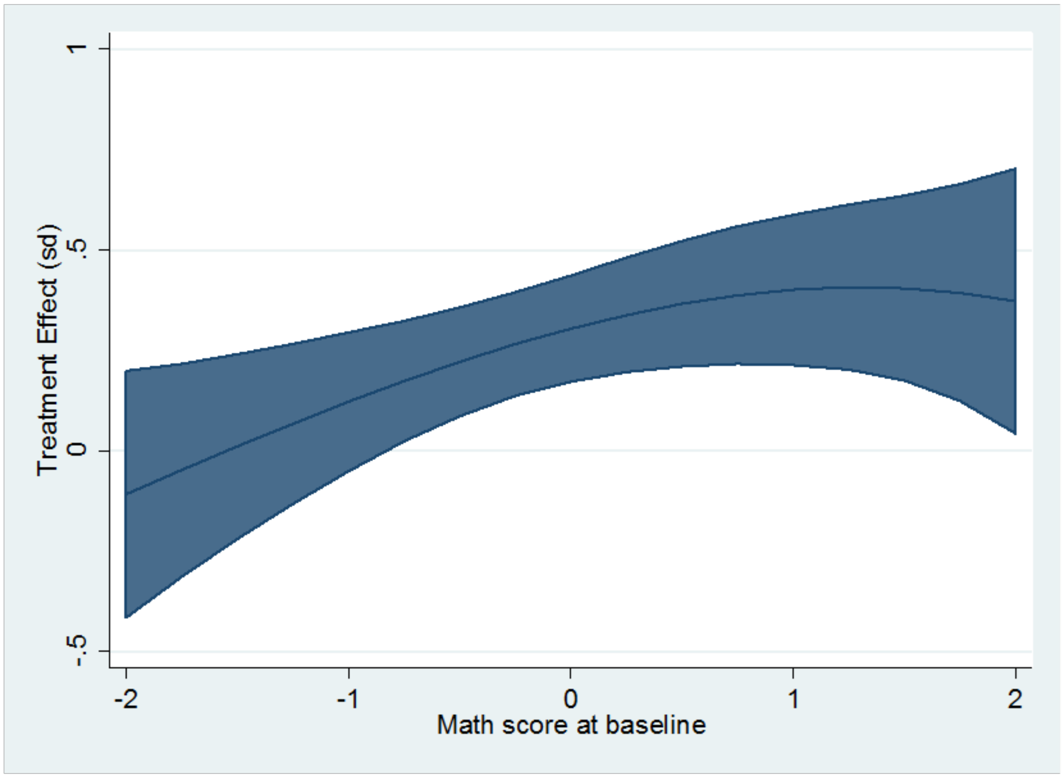
	Math		Reading (Georgian as native language)	
	Male	Female	Male	Female
G-PriEd	0.24***	0.30***	0.13	0.18*
	(0.071)	(0.075)	(0.066)	(0.078)
Baseline math	0.47***	0.39***	0.23***	0.19***
	(0.044)	(0.049)	(0.037)	(0.039)
Baseline reading	0.33***	0.39***	0.43***	0.49***
	(0.040)	(0.045)	(0.035)	(0.042)
Obs	880	816	770	702

Note: All specifications include students' age and special education status, categorized student-teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for 11 regions. Standard errors clustered at the school level in parentheses.

Source: G-PriEd and EMIS data for 2013 and 2015.

Figures

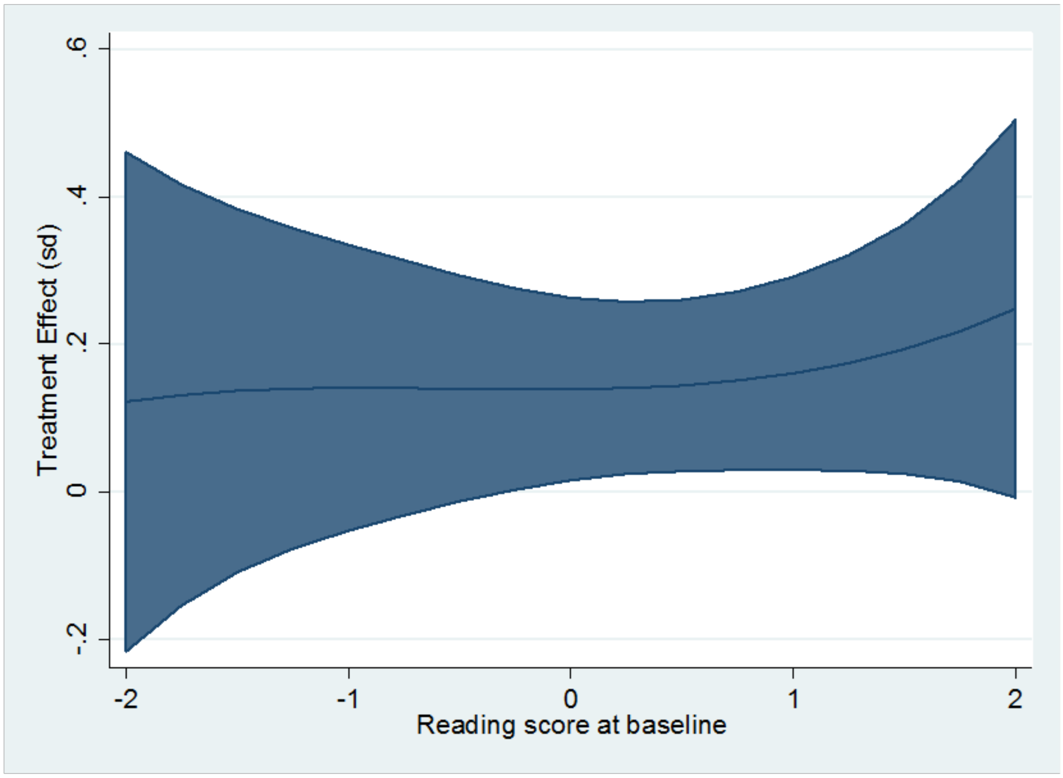
Figure 1. Effect of G-PriEd on math by baseline test scores with 95% CIs



Note: Regression pools students in grades 3-6 at endline, and includes students' age, gender and special education status, categorized student teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for language of test, 11 regions and cohort fixed-effects. Standard errors calculated using the delta method. The distribution of baseline test scores is trimmed so only values between -2 and 2 are shown.

Source: G-PriEd and EMIS data for 2013 and 2015.

Figure 2. Effect of G-PriEd on reading (Georgian as native language) by baseline test scores with 95% CIs



Note: Regression pools students in grades 3-6 at endline, and includes students' age, gender and special education status, categorized student-teacher ratio, fraction of certified teachers, class size and school size at baseline, and dummies for 11 regions and cohort fixed-effects. Standard errors calculated using the delta method. The distribution of baseline test scores is trimmed so only values between -2 and 2 are shown.

Source: G-PriEd and EMIS data for 2013 and 2015.