IDB

ove Office of Evaluation and Oversight

The Pedagogy of Science and Environment: Experimental Evidence from Peru

Working paper

September 15, 2012

OVE/WP-01/12

IDB

ove Office of Evaluation and Oversight

The Pedagogy of Science and Environment: Experimental Evidence from Peru

Diether Beuermann
Emma Naslund-Hadley
Inder J. Ruprah
Jennelle Thompson

September 15, 2012

© Inter-American Development Bank,
www.iadb.org

The information and opinions presented in this publication are entirely those of the author(s), and no endorsement by the Inter-American Development Bank, its Board of Executive Directors, or the countries they represent is expressed or implied. This paper may be freely reproduced provided credit is given to the Inter-American Development Bank.

The Pedagogy of Science and Environment: Experimental Evidence from Peru.

Abstract
In today's knowledge-based societies, understanding basic scientific concepts and the capacity to structure and solve scientific questions is more critical than ever. Accordingly, in this paper we test an innovative methodology for teaching science and environment in public primary schools where traditional (teacher centred) teaching was replaced with student centred activities using LEGO kits. We document positive and significant improvements of 0.18 standard deviations in standardised test scores. Such positive results are mainly concentrated within boys that were located above the median of baseline academic performance.

Keywords: science; environment; Peru

JEL: I21; I28; I29; O15; O31

1) Introduction

School enrolment has increased significantly in the last decade or so in most developing countries, but the quality of education has not progressed as rapidly. Evidence suggests that increased school participation does not automatically translate into increased competency in basic skills (Glewwe and Kremer, 2006; Glewwe et al, 2011). Accordingly, Peru is not an exception. The coverage rates of initial, primary and secondary education have increased significantly in the last decade and are in general considered satisfactory for a developing, middle income country.[i]

However, in terms of quality, in national, regional and international tests, Peruvian students repeatedly score poorly. In 2009, a national test of second graders revealed that only 13.8 per cent achieved the expected learning outcomes in mathematics.[ii] In the Second Regional Comparative Education Study (SERCE), Peruvian students scored below the regional average in both mathematics and natural science (UNESCO, 2008). In third grade mathematics, more than half of the student population reached only the very

lowest achievement level. A further finding of the SERCE evaluation was that urban students' average test scores were three times greater than those of rural pupils. In the 2009 Program for International Student Assessment (PISA), among the 65 participating countries, Peru ranked 60 in mathematics and 63 in natural sciences (OECD, 2011). This test which has as its main objective to evaluate to what degree students who are about to finish secondary school (15 year olds) have acquired the necessary skills to fully participate in the knowledge society, suggests that the majority of Peruvian students are neither prepared to enter the labour market nor for initiating tertiary studies. Thus a critical policy issue is how to improve education quality in Peru.

There is a growing body of research that supports a shift from traditional teacher led to student centred learning complemented with some degree of inquiry as a mean to maximise learning (Healy, 1990; Lowery, 1998). Recently, through a meta-analysis of 37 experimental and non-experimental studies of inquiry-based instruction, Furtak et al (2012) found that teacher led inquiry-based approaches were more effective than pure student led approaches. Therefore, additional research is required to define what degree of inquiry is most effective for teaching different subjects and contents. Rigorous evaluations of natural sciences teaching approaches in developing countries are particularly scarce. Against this background, in 2008 the Peruvian Ministry of Education (MOE) requested assistance from the Inter-American Development Bank to develop and validate a student centred pedagogical approach for science and environment education with a complement on inquiry. This partnership resulted in the development of a guided inquiry approach (Colburn, 2000) for the natural science classroom, focusing on student centred activities under teacher guidance.

Rather than teaching students to simply memorize the history of science and scientific facts (as it is traditionally done), the new approach focuses on the development of scientific thinking and an understanding of what they can do with their knowledge. The methodology builds on children's curiosity and natural proclivity to explore the world around them. The new methodology encompasses three modules – our environment, the human body, and our physical world – which were piloted in third grade classes within 53 treated schools in the department of Lima and evaluated via an experimental design (with 53 comparison schools that continued with the traditional teaching approach).

The treatment consisted in the development of didactic materials, teacher training modules and classroom support, as well as a continuous student assessment instruments. In terms of equipment, classroom laboratory equipment was provided to support experiments both in and outside of the classroom, including LEGO Data learning materials. While LEGO use has previously been evaluated, most of the early studies consisted of small sample sizes and limited study periods and are best characterised as qualitative rather than rigorous evaluations. Nonetheless, such qualitative evaluations found that use of these materials increased critical thinking, developed abilities in problem solving and enhanced collaboration between pupils (Noble, 2001). In Peru, Iturrizaga (2000) found that test scores improved significantly in mathematics, reading, technology and eye-hand coordination using data from treatment and comparison groups, although not randomly selected. In addition, Hussain et al (2006), using Swedish data, found that treated fifth grade students performed better in mathematics with this pedagogical approach. The limited number and methodological weakness of the studies

suggests further research is needed on the causal impact of the pedagogical approach on student's learning.

In this paper we report the findings of an experimentally designed evaluation of the Peruvian pilot, which draws on standardised tests administered in the treatment and control groups as well as from surveys of principals, teachers, students and parents. The rest of this paper proceeds as follows: Section two describes the program background, the learning areas covered and the context in which it was implemented. Section three presents the research strategy and its implementation plus the quality of the data. Section four presents the results and their interpretation. Finally, section five concludes.

2) The programme

2.1. Background

Before the mid twentieth century, natural science curriculum design assumed that a child was a blank slate when entering the education system without notions or beliefs about different facts and phenomena. The teacher therefore was tasked with transferring his or her knowledge of scientific concepts to the students. This frontal teaching style was later replaced in the literature by a range of the active learning approaches, including role-playing, student debates, and collaborative learning groups.

In Latin America, the debate over teacher led approaches versus student centred learning remains on the philosophical and ideological levels. Although some education systems have updated learning plans, in practice the full frontal teaching style continues to dominate classrooms throughout the region. Peru is no exception with science curricula across grade levels, emphasizing inquiry based learning while classrooms are characterized by traditional

chalk and talk teaching styles that prioritise the memorisation of scientific concepts and the history of science over the development of abilities of inquiry and critical thinking.

In third grade, the traditional science and environment curriculum encompasses three thematic areas. First, the "human body" covers the structures, functions and interactions of the different systems of the human body. Second, "our environment" introduces students to scientific explanations about how ecosystems work. In terms of content knowledge, it covers three topics: (i) ecosystems, including how plants develop and grow, and livening things (organisms, animal and plant link, vertebrate and invertebrate animals); (ii) biodiversity, including native and exotic animals and plants; and (iii) protection of plants, animals and habitats. Third, a module called "our physical world" covers topics such as the planet earth and its characteristics, forces and movement, electricity, light and colour, and magnetism (Ministry of Education, 2008).

2.2. The intervention

The intervention developed student centred methodologies complemented with inquiry to teach the same areas and topics covered under the traditional pure teacher led approach. Under the developed guided inquiry protocol, the teacher challenges the students by providing the problem to be solved as well as the materials. The students are expected to elaborate their own procedures, record and report their results. The teacher facilitates learning by motivating students to explore new ideas and formulate interesting questions. During the conversations, the teacher introduces the formal names of different concepts. The students are then expected to apply the concepts to new situations.

The previous design drew on several active learning approaches (Bonwell and Eison, 1991) that emphasizes the importance of understanding science rather than memorization of isolated concepts. This included building new knowledge on what students already know; developing critical thinking skills; teaching through inquiry, and addressing different student learning styles.

Under the "human body" area, the model aimed to stimulate curiosity and draw logical conclusions about the topics covered. In "our environment" students were introduced to field work, helping learners generate their own scientific explanations about how ecosystems work based on empirical evidence rather than only presenting them in class. In the "physical world" area, the program aimed for students to become more proficient in carefully formulating research questions, conducting experiments and interpreting data related to the topics covered.

To do so, several materials needed to cover the curriculum standards of the three modules were developed. These included lesson plans, activity journals and simple classroom kits with microscopes, magnifying glasses, measuring cups, scales and consumable materials to teach the "environment" and "human body" modules, as well as LEGO educational kits to teach the "physical world" module.

The considerable pedagogical and content gaps of Peruvian teachers constituted a challenge for successfully bringing the new inquiry based curriculum to the science classrooms. A cornerstone of the program therefore was to help teachers both develop adequate knowledge of third grade science content, and learn to support student learning through inquiry. The teacher training encompassed two types of activities: (i) interactive workshops to develop content knowledge, suggest a range of methods that elicit and challenge students' thinking,

and help teachers develop classroom activities that allow learners to engage in scientific investigations; and (ii) technical assistance and teacher tutoring inside and outside the classroom, including demonstration sessions, tandem classes, and individual feedback.

In addition to the development of content and pedagogical knowledge, the workshops and one to one assistance emphasised the development of formative student assessment to help teachers develop individual learning plans for students. All teacher training and tutoring sessions aimed to be as concrete and hands on as possible, prioritizing step by step classroom activities over abstract philosophical conversations about definitions of inquiry based learning.

2.3. The context

To grasp an understanding of the context in which the program took place, we draw on the data collected on household socio demographics. We gathered this information from the pilot schools and from the students' families via surveys.

About one third of the students surveyed report that they work in the local family farm or in the family's microenterprise at some moment during the school year. The percentage of students who reported that they sell products in streets is about 8 per cent to 10 per cent depending upon the season of the year. Surprisingly, more rural students report working in the streets than urban students, including those living in the Lima metropolitan area.

For example, 11 per cent of rural students report selling products in the streets in weekends during the school year, while only 6 per cent of urban students report doing so. The survey of parents reveal that 46 per cent of household heads had not finished secondary education and only 19 per cent report having some education beyond

secondary school. Surveyed parents report that their average income is S/500.00 (about US$180) per month. This is below the minimum wage that was in place during the studied period.[iii] Although there are a number of reasons to doubt the exact income figure it is clear that the majority of the students are from low income families.

3) Research strategy: design, implementation and data

3.1. Conceptual framework

The typical conceptual framework invoked in the analysis of education is the production function approach; a conventional time varying linear specification of which is:[iv]

$$A_t = \beta_0 + \beta_{1t} \cdot S_t + Q_t' \beta_{qt} + C_t' \beta_{ct} + H_t' \beta_{ht} + I_t' \beta_{It} + \mu_{at} \quad (1)$$

where A_t is skills learned at time t, S_t is years of schooling acquired by time t, Q_t is a vector of school and teacher characteristics, C_t is a vector of child characteristics, H_t is a vector of household characteristics and I_t is a vector of school inputs under parental control, and u_{at} is an error term. Where u_{at} accounts for variables for which there is no data and measurement errors in skills learnt or in the explanatory variables in the equation. The causal impact of a particular explanatory variable on skills can be consistently estimated only if u_{at} is uncorrelated with such explanatory variable. This situation is unlikely to hold in observational data due to omitted variable bias, selection, attrition, and measurement error amongst other factors.

To overcome these potential biases and recover the causal effect of the treatment, we use data generated from a randomised intervention. This consists of introducing a simple change to a set of randomly selected schools and not implementing it in other randomly selected group of schools. Well done randomised trials can overcome the problems

discussed above to uncover causal effects as the difference in the outcome of interest, in our case test scores, between treated and non-treated groups. The best way to use experiments to inform policy is to test policies, further, as argued by Glewwe et al (2011); this approach has much promise where theory provides little guidance such as what types of pedagogical materials are the most effective.

Therefore, our focus is to provide consistent evidence regarding the effectiveness of a particular teaching methodology rather than estimating a particular form of education production function or determining relative weights between different components of such function. In short, we exogenously affect one particular input in order to assess its effects on academic performance holding constant all other observable and unobservable inputs that might be present in the unknown education production function.[v] For this purpose, a randomised intervention is a robust approach. The literature on educational program evaluation has used this methodology extensively to analyze learning effects of different educational inputs such as computers (Barrera-Osorio and Linden, 2009; Malamud and Pop-Eleches, 2011; Cristia et al, 2012), class size and tracking (Duflo et al, 2011), and educational software (Banerjee et al, 2007; Carrillo et al, 2010).

3.2. Design

The research strategy consisted of three components. First, an experimental designed evaluation to estimate the causal impact of the new pedagogical approach on scholastic achievement of the students. Second, surveys of principals, teachers, parents and students were carried out to obtain the socio-demographic information of the schools and the

students' families. Third, a qualitative evaluation was carried out to further understand the context of the treatment and the evaluation.

The experimental design covered 106 schools in the Department of Lima where the random assignment of the treatment was at the school level, that is, 53 treated and 53 non-treated schools, with a total of 2,771 third grade students in the 106 schools. The sample was stratified according to school location (urban, metropolitan, and rural). For budgetary reasons only two classrooms per school were included. In the schools where there were more than two classes, the two classes included in the evaluation were chosen randomly in the presence of the principal. Previous studies suggest that there is a high correlation between classes in the same school hence little additional information is lost by only including two classes per school.[vi] Baseline exams were applied in Science and Environment as well as Mathematics and Reading Comprehension to students starting third grade and again at the end of the year (2010 academic year running from April through December).[vii]

Surveys were applied to school principals, teachers, and the students' parents. The survey of the schools' principals aimed to obtain information on the number of students and teachers, the school's facilities, equipment and didactic materials and the school's climate. The survey of teachers collected baseline data on diverse aspects of teaching science and the environment in their school and to their students. A socio-demographic parent survey was given to the students to be brought home, answered by their parents and returned in closed envelopes.

The quantitative research tools were complemented by a qualitative evaluation. The qualitative evaluation consisted of visits in situ to eight schools, four from the treated and

four from the non treated groups. The eight schools were chosen such that the strata of the study; rural, urban, metropolitan, small and medium sized schools, were covered. The schools were visited during August and September of 2010. This additional analysis included interviews with teachers, principals, and third grade pupils. The aim was to convey additional information on the context of the experiment, how that context could affect the experiment, hence, the factors that contributed to the successes and failures of the experiment. We use the observations gathered during this stage when analyzing our quantitative results in order to interpret them.

3.3. Implementation

The implementation of the experiment deviated from that which had been planned in a number of ways. First, the intervention was planned to be administered during the entire academic year 2010 (April through December). However, delays in the distribution of materials and training sessions determined that the treatment was only implemented during five effective months before evaluating the students (July to November).

Second, the teaching materials for the modules "human body" and "environment" were unable to be distributed due to logistical complications in their import process. Only the LEGO kits for the "physical world" module were effectively used during the five month implementation period. Finally, the training time of rural teachers was less than those in urban schools (20 hours compared to 60 hours). This because the web based teaching support system was not available to rural schools given their lack of internet connection. In addition, rural training was executed later determining that rural schools had only two months of post training implementation. Altogether these issues implied that the duration

of treatment was less than that planned and that the intensity and duration of treatment was even lower in rural schools relative to urban schools.

3.4. Data

Properly executed random assignment should eliminate, on average, all potential confounders both observable and non observables. Table 1 presents summary statistics of the physical characteristics of the schools and the exam results at baseline by treatment status.

Table 1: Differences between Treated and Control Groups at Baseline

	Treatment (1)	Control (2)	Difference (3)	Observations (4)
Panel A: School Characteristics				
Number of teachers - Primary	12.31	10.84	1.47	106
	(1.60)	(1.29)	(2.05)	
Connected to piped water	0.73	0.70	0.03	106
	(0.06)	(0.07)	(0.09)	
Has telephone	0.48	0.60	-0.12	106
	(0.07)	(0.07)	(0.10)	
Connected to internet	0.31	0.22	0.09	106
	(0.07)	(0.06)	(0.09)	
Plot for growing plants	0.94	0.98	-0.04	106
	(0.03)	(0.02)	(0.04)	
Computer room	0.87	0.86	0.01	106
	(0.05)	(0.05)	(0.07)	
Science room	0.85	0.96	-0.11*	106
	(0.05)	(0.03)	(0.06)	
Art room	0.91	0.98	-0.08	106
	(0.04)	(0.02)	(0.05)	
Music room	0.91	0.98	-0.08	106
	(0.04)	(0.02)	(0.05)	
Panel B: Performance (all schools)				
Science and Environment	0.12	0.00	0.12	2771
	(0.08)	(0.08)	(0.11)	
Verbal ability	0.12	0.00	0.12	2771
	(0.06)	(0.07)	(0.09)	
Mathematical ability	0.07	0.00	0.07	2771
	(0.06)	(0.06)	(0.09)	

Table 1: Differences between Treated and Control Groups at Baseline

	Treatment (1)	Control (2)	Difference (3)	Observations (4)
Panel C: Performance (schools with 1 or 2 sections)				
Science and Environment	-0.06	-0.06	0.00	1487
	(0.11)	(0.08)	(0.13)	
Verbal ability	-0.04	-0.10	0.06	1487
	(0.09)	(0.09)	(0.13)	
Mathematical ability	-0.04	-0.11	0.07	1487
	(0.08)	(0.09)	(0.12)	
Panel D: Performance (schools with 3 or more sections)				
Science and Environment	0.32	0.07	0.25	1284
	(0.09)	(0.14)	(0.17)	
Verbal ability	0.30	0.12	0.18*	1284
	(0.05)	(0.08)	(0.09)	
Mathematical ability	0.19	0.13	0.06	1284
	(0.09)	(0.08)	(0.12)	

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Test scores are expressed in standard deviations with respect to the control group.

There are no significant differences in school infrastructure between treated and control groups (Panel A). Similarly there are no statistical differences in the test scores of students in the treated and control schools (Panel B).[viii] Given that our research strategy sampled only two random sections within schools with three or more third grade sections, is important to show that such procedure did not contaminated or potentially biased our design. Therefore, we compare treated and control schools with respect to baseline grades according to the number of third grade sections. Panel C shows that there were no significant differences between treated and control schools with one or two sections. Similarly, Panel D shows that baseline differences were inexistent when considering schools with three or more sections. In that way, we can confidently use our sample without worrying regarding potential biases that might have been introduced by differential non random section sampling within schools with three or more sections.

In addition, baseline test scores provide rich information to test for potential heterogeneous impacts across the initial test scores' distribution. Thus an important evaluative question is whether the new pedagogical model can help closing learning gaps. To address this question we will estimate intervention effects for subgroups. The potential heterogeneity of impacts and whether they accentuate or attenuate the original test score inequalities is a critical policy design issue regarding potential uniform against differentiated expansion of the program.

The relatively high attrition is a potential problem for the evaluation. The end line sample of the number of students that took the exams fell by 14.4 per cent to 2,373 students from 2,771 students in the baseline.[ix] This attrition of the original sample could result in bias if the students that dropped out are systematically different from those that remain. The attrition bias would undermine the internal validity of the study and the reduced sample could reduce the statistical power of the tests. Attribution compromises the internal validity of the experimental design because treated and control group members for whom follow-up data are available may be non-random sub-samples of the original groups. Therefore, below we present the results of testing if the intervention led to drop outs, if the attrition was orthogonal to the outcomes of interest and if the power of the sample was adequate.

To test if attrition was not systematically related to the treatment, the following regression was estimated:

$$L_{ij} = \delta + \beta \cdot T_j + \varepsilon_{ij} \quad (2)$$

Where L_{ij} is equal to unity if the student *i* in school *j* was not evaluated in end line and zero otherwise ("Leaver"); T_j is equal to unity if the school *j* was treated and zero

otherwise. The parameter β will be statistically indistinguishable from zero if attrition rates were not systematically different between the treated and control groups.

Column one of Table 2 reports the β estimate from equation (2). The estimated coefficient for β is statistically indistinguishable from zero showing that attrition rates in the treated and control groups were not systematically different.

Table 2: Attrition Tests for the Test Scores

Dependent Variables:		Baseline Standardised Scores		
	Leaver	Science & Environment	Reading	Math
	(1)	(2)	(3)	(4)
Treated	-0.03	0.11	0.11	0.06
	(0.02)	(0.11)	(0.09)	(0.09)
Treated x Leaver		-0.01	0.02	0.05
		(0.15)	(0.14)	(0.12)
Observations	2771	2771	2771	2771

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Test scores are expressed in standard deviations with respect to the control group.

The second test's hypothesis is that attrition was orthogonal to the outcomes of interest. To empirically determine if this was so the following regression was estimated:

$$Y_{ij} = \delta + \delta_1 \cdot L_{ij} + \beta_1 \cdot T_j + \beta_2 \cdot T_j \cdot L_{ij} + \varepsilon_{ij} \quad (3)$$

Where Y_{ij} is the outcome of interest (standardised test scores at baseline for our case) for student *i* in school *j*. The other variables are defined as in equation (2);[x] β_1 captures the difference at baseline of treated and control students that were tested at both baseline and end line; while (β_1 + β_2) measures the difference, also at baseline, between treated and non-treated students that were only tested at baseline. For a causal interpretation of the impact estimations both coefficients should be statistically indistinguishable from zero. As shown in the columns two, three, and four of Table 2; parameters β1 and β_2 are

statistically insignificant for the three topics; science and environment, reading comprehension and mathematics.

Attrition could have reduced the power of the initial experimental design. Fortunately, this was not the case as the minimum detectable effects (MDEs) are exactly the same (0.19 standard deviations) despite the reduction in sample size at the pupil level. This is explained because treatment assignment was done at the school level and the power of the experiment is, therefore, primarily driven by the number of schools involved (which was the same between baseline and end line).

4) Results and interpretation

4.1. Overall effects

The direct impact expected from the program is an improvement in the test scores of science and environment. To measure the impact of the program on the test scores the following regression was estimated:

$$Y_{ij,t} = \delta + \delta_1 \cdot Y_{ij,t-1} + \beta \cdot T_j + \varepsilon_{ij,t} \quad (4)$$

Where $Y_{ij,t}$ is the standardised, at end line, exam result for student *i* in school *j*; $Y_{ij,t-1}$ is the exam result at baseline.[xi] The parameter β measures the impact of the program (expressed in terms of standard deviations with respect to the control group) on the exam results. The results of the estimated regression are given in Table 3.

Table 3: Overall Results - Science and the Environment

Dependent Variables:	Science & Environment							
	Human Body		Environment		Physical World		Overall	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Overall Effects								
Treatment effect	0.02	-0.03	0.08	0.03	0.22**	0.18**	0.15	0.10
	(0.09)	(0.06)	(0.10)	(0.07)	(0.10)	(0.08)	(0.11)	(0.07)
Observations	2536	2373	2536	2373	2536	2373	2536	2373
Panel B: Effects by Gender								
Girls	-0.06	-0.06	-0.03	-0.03	0.10	0.10	0.02	0.02
	(0.10)	(0.07)	(0.11)	(0.07)	(0.11)	(0.09)	(0.12)	(0.08)
Additional effect on Boys	0.14	0.06	0.20**	0.11	0.24***	0.16**	0.25***	0.15**
	(0.09)	(0.08)	(0.08)	(0.07)	(0.08)	(0.08)	(0.09)	(0.07)
Observations	2536	2373	2536	2373	2536	2373	2536	2373
Panel C: Geographical Effects								
Rural	-0.23	-0.18	-0.20	-0.16	0.02	0.09	-0.14	-0.08
	(0.14)	(0.12)	(0.18)	(0.13)	(0.16)	(0.13)	(0.18)	(0.13)
Urban	0.07	0.01	0.16	0.10	0.26**	0.20**	0.21	0.14*
	(0.10)	(0.07)	(0.12)	(0.07)	(0.12)	(0.09)	(0.13)	(0.08)
Observations	2536	2373	2536	2373	2536	2373	2536	2373
Panel D: Geographical Effects by Gender								
Rural - Girls	-0.29	-0.21	-0.30	-0.23*	-0.10	-0.03	-0.26	-0.18
	(0.18)	(0.14)	(0.19)	(0.13)	(0.18)	(0.15)	(0.20)	(0.14)
Rural - Boys	-0.17	-0.15	-0.10	-0.09	0.14	0.21	-0.02	0.02
	(0.16)	(0.15)	(0.20)	(0.17)	(0.17)	(0.14)	(0.20)	(0.17)
Urban - Girls	-0.00	-0.02	0.06	0.06	0.15	0.14	0.10	0.08
	(0.11)	(0.07)	(0.12)	(0.08)	(0.14)	(0.11)	(0.14)	(0.09)
Urban - Boys	0.13	0.04	0.25*	0.15	0.36***	0.26**	0.32**	0.20**
	(0.11)	(0.09)	(0.14)	(0.09)	(0.13)	(0.10)	(0.14)	(0.10)
Observations	2536	2373	2536	2373	2536	2373	2536	2373
Baseline controls	No	Yes	No	Yes	No	Yes	No	Yes

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. All scores are expressed in standard deviations with respect to the control group. Estimations in columns one, three, five, and seven are obtained by estimating the model without controlling by the initial test scores. Therefore, we used all the observations, 2,536, at the end line thus included pupils that were not tested at the base line. The reported estimations in columns two, four, six, and eight are controlled for initial exam results of each pupil therefore includes pupils that were tested at the base line and the end line, that is, 2,373 pupils.

As shown in Panel A of Table 3 there are no impacts on the Human Body and Environment modules (columns one to four) and the overall test score (columns seven and eight). However, there is a significant effect on the Physical World module, with an estimated impact of 0.18 standard deviations (column six). The results show that this new pedagogical method is more effective than those traditionally used in Peruvian schools;

despite the fact that the application of the new method was for less time than planned. Further, the size of the effect is higher than that found in the majority of studies on interventions aimed at improving test scores at the school level (see Glewwe et al, 2011).

Only finding effects in the physical world section is not surprising given that the didactic materials for the other sections were not used effectively; while the LEGO kits were indeed used. Furthermore, qualitative evidence corroborates that the program generated interest in pupils regarding this module. Indeed, teachers pointed out that school assistance had always been a problem. However, when the program began the days the LEGO kits were to be used was preannounced. Pupils who were often absent began to come more regularly on those days. Then teachers stopped pre announcing the days that the LEGO kits were to be used resulting in an increase in assistance for all the days. Teachers also reported how children started to identify machines that their parents used at work and understanding scientific concept such as pulleys.

It is important to determine if there were differential impacts for distinct segments of the targeted population. Absent any heterogeneity in effects, the Physical World module can be expanded uniformly but with heterogeneous effects any expansion will require modifications to the pedagogical approach used for different sub groups. We consider the possible differential impacts by gender, geographical location of the school and baseline performance. In addition we attempt to determine if there were spillover effects of the treatment on mathematics and reading comprehension.

4.2. Impacts by gender

An important aspect is to determine if the treatment had differential impacts by gender. Thus, Table three – Panel B shows the impacts by gender. To obtain the differential effects by gender the following regression was estimated:

$$Y_{ij,t} = \delta + \delta_1 \cdot Y_{ij,t-1} + \delta_2 \cdot Male_{ij} + \beta_1 \cdot T_j + \beta_2 \cdot T_j \cdot Male_{ij} + \varepsilon_{ij,t} \quad (5)$$

Where the categorical variable $Male_{ij}$ equals unity if student *i* in school *j* is a male. Parameter β_1 captures the impacts of the program on girls; while β_2 captures the additional program impacts on boys with respect to girls.

Consistent with the previous results there are null effects for the sub topics Human Body and Environment. However, the impacts on the Physical World module is exclusively on boys (columns five and six), that is, an additional impact of 0.24 to 0.16 standard deviations, while for girls the impact is statistically indistinguishable from zero. Further, the program's impact on boys overall test scores in Science and the Environment is significantly higher than girls by 0.15 standard deviations while there was no impact on girls' overall test scores (column eight).

Thus the treatment accentuated gender inequality. A number of possible factors could have contributed to this differential effect. The gender difference may be the result of gender differences in the appeal of the specific content areas covered within the three modules, including the specific science problems that students were asked to solve. Given the limited number of science equipment and LEGO teaching kits per classroom (only one kit to be shared within the class), the kits may have been monopolized by boys. While no systematic data on the use of the teaching materials was recorded; on-site visits for the qualitative interviews suggest that indeed boys' monopolization was the case.

4.3. Impacts by geographical location

In addition, the data allows determining whether there are differential impacts by geographical location of the school. The random assignment was stratified geographically such that of the 53 (53) treatment (control) schools, 29 (28) were in urban areas and 24 (25) were in rural areas. Panel C of Table three shows the estimated impacts by geographical location.

Panel C provides estimated impacts within rural and urban schools separately. Column six reveals a significant effect in urban areas equivalent to 0.2 standard deviations but no effect in rural areas for the Physical World module. We find no significant effect in any of the other individual modules but we do find some effect in the total test score of 0.14 standard deviations (column eight). These differential effects reflect the fact that the pilot had a more intensive intervention within urban areas. The training of rural teachers was less intense than urban teachers (20 hours compared to 60 hours). In addition, the training of rural teachers was implemented later than urban teachers such that the effective period with the new teaching method was less than two months in rural areas before the application of the final exam.

Therefore, given that the program was implemented better in urban areas, is of much interest to assess whether its effects varied by gender within these geographical areas. For instance, Panel D presents these differential effects. First, we see no differential effects within rural areas as virtually all estimated impacts are statistically indistinguishable from zero. Within urban areas we don't see any effects for girls. However, effects for boys appear to be significant in physical world and the overall score (columns five to eight). The overall effect for the test amounts to 0.2 standard deviations (column eight). This

finding provides evidence that the program with adequate training for teachers (60 hours) and at least 5 months of application provides positive learning outcomes in overall performance for the segment that appears to monopolize the didactic LEGO kits (that is boys).

4.4. Distribution of the impacts

The treatment may wide or tighten the distribution of test scores. Panel A of Table four shows differential impacts by quartiles of the baseline exam results. That is, students were ranked according to their baseline test scores and the distribution was categorised into quartiles. The following regressions were then estimated:

$$Y_{ij,t} = \delta + \delta_1 \cdot Y_{ij,t-1} + \delta_2 \cdot Q1_{ij,t-1} + \delta_3 \cdot Q2_{ij,t-1} + \delta_4 \cdot Q3_{ij,t-1} + \beta_1 \cdot T_j \cdot Q1_{ij,t-1} + \beta_2 \cdot T_j \cdot Q2_{ij,t-1} + \beta_3 \cdot T_j \cdot Q3_{ij,t-1} + \beta_4 \cdot T_j \cdot Q4_{ij,t-1} + \varepsilon_{ij,t} \quad (6)$$

Where $Q_{1ij,t-1}$, is equal to unity if the student *i* in school *j* was between the first and 25th percentile of the base line test results distribution while zero otherwise; $Q_{2ij,t-1}$ is equal to unity if the student *i* in school *j* was between the 26th and 50^{th} percentile in the base line test results while zero otherwise; $Q_{3ij,t-1}$ is equal to unity if the student *i* in school *j* was between the 51 and 75 percentile of the baseline distribution while zero otherwise; $Q_{4ij,t-1}$ is equal to unity if the student *i* in school *j* was above the 75th percentile while zero otherwise. The rest of variables are defined as before. In this setting, the estimates of parameters β_1, β_2, β_3, and β_4 represent the impacts of the program within each quartile of the baseline distribution.

Table 4: Results by Baseline Test Scores - Entire Sample

Dependent Variables:	Science & Environment			
	Human Body	Environment	Physical World	Overall
	(1)	(2)	(3)	(4)
Panel A: Quartile Effects - Entire Sample				
Quartile 1	-0.09	0.04	0.09	0.03
	(0.09)	(0.08)	(0.10)	(0.09)
Quartile 2	-0.06	0.09	0.11	0.07
	(0.11)	(0.08)	(0.10)	(0.10)
Quartile 3	0.00	-0.09	0.27***	0.11
	(0.08)	(0.09)	(0.10)	(0.09)
Quartile 4	0.01	0.03	0.25**	0.15*
	(0.06)	(0.08)	(0.10)	(0.08)
Observations	2373	2373	2373	2373
Panel B: Quartile Effects - All Girls				
Quartile 1	-0.14	-0.01	0.08	-0.01
	(0.13)	(0.12)	(0.13)	(0.13)
Quartile 2	-0.08	-0.02	0.04	-0.02
	(0.13)	(0.11)	(0.14)	(0.12)
Quartile 3	0.01	-0.14	0.15	0.03
	(0.10)	(0.10)	(0.11)	(0.10)
Quartile 4	-0.04	0.03	0.10	0.05
	(0.10)	(0.10)	(0.14)	(0.11)
Panel C: Quartile Effects - All Boys				
Quartile 1	-0.04	0.09	0.10	0.07
	(0.12)	(0.10)	(0.11)	(0.11)
Quartile 2	-0.03	0.20*	0.17	0.14
	(0.13)	(0.10)	(0.10)	(0.10)
Quartile 3	0.00	-0.05	0.37***	0.18
	(0.11)	(0.11)	(0.12)	(0.12)
Quartile 4	0.06	0.03	0.40***	0.24**
	(0.09)	(0.11)	(0.11)	(0.10)
Observations	2373	2373	2373	2373

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Test scores expressed in standard deviations with respect to the control group.

Similar to the global results there is no impact for the Human Body and Environment modules (columns one and two). However, the impacts on the Physical World module are concentrated for students who were in the top half of the baseline test score distribution (column three). Thus the intervention accentuates the distribution where students with higher scores at the time of the baseline measurement benefit more than students with

relatively lower ex-ante scores. Furthermore, it appears to be some effect in overall test scores but only for the top quartile (column four). This result has important policy design ramifications. One possible policy design option would be to avoid accentuating the inequality of impacts of the intervention by sorting students according to their initial preparedness (baseline test scores); an approach often called tracking. As Duflo et al (2011) study shows that when students are divided into groups with similar ability, as measured by baseline test scores, then student's performance improves across the entire distribution and not just for the top scorers. The sorting allows the speed of teaching to be closer aligned with the needs of the individual students.

However, it could also be the case that only high ability students are able to incorporate the skills of the intervention. Therefore, even with tracking, low ability students would not take advantage of the program. In such scenario, would be appropriate to still use the program for high ability students perhaps with a tracking design. But an alternative method would need to be developed for low achievers. Unfortunately, our intervention did not incorporate any tracking treatment and therefore we are unable to illuminate this issue that is left for future research.

4.5. Results by gender, location and baseline test score's distribution

Previously we had noted that the positive impacts were located in the top half of the baseline distribution of test scores and that there was no overall impact on girls test scores. In Panels B and C of Table four we look at impacts by gender and the initial distribution of the scores.

First, there is no differential impact for girls on test scores, either in sub-topics or overall (Panel B). According to the distribution of test scores at baseline; the impact is statistically indistinguishable from zero throughout the distribution. For boys (Panel C) there is a positive effect for the Physical World module and overall test scores (columns three and four). We also notice that the size of the effects is increasing in baseline scores. While the bottom half had (insignificant) impacts ranging from 0.1 to 0.17 standard deviations; the top half had significant impacts ranging from 0.37 to 0.4 standard deviations (column three). A similar pattern arises for overall test scores where the effects range from an insignificant impact of 0.07 standard deviations within the bottom quartile to a significant impact of 0.24 standard deviations at the top quartile (column four).

As we have previously found that the main effects were observed in urban areas, we repeat the analysis of heterogeneous effects according to baseline scores and gender only for urban schools.[xii] Table 5 below reports the results.

Table 5: Results by Baseline Test Scores - Urban Sample

Dependent Variables:	Science & Environment			
	Human Body	Environment	Physical World	Overall
	(1)	(2)	(3)	(4)
Panel A: Quartile Effects - Entire Urban Sample				
Quartile 1	-0.08	0.14	0.13	0.09
	(0.10)	(0.09)	(0.12)	(0.11)
Quartile 2	0.01	0.19**	0.12	0.13
	(0.11)	(0.09)	(0.13)	(0.12)
Quartile 3	0.08	-0.01	0.29**	0.17
	(0.08)	(0.09)	(0.13)	(0.11)
Quartile 4	0.00	0.03	0.23**	0.13
	(0.06)	(0.09)	(0.12)	(0.09)
Observations	1718	1718	1718	1718
Panel B: Quartile Effects - Urban Girls				
Quartile 1	-0.02	0.08	0.14	0.09
	(0.15)	(0.13)	(0.16)	(0.15)
Quartile 2	-0.07	0.04	0.09	0.04
	(0.15)	(0.12)	(0.18)	(0.15)
Quartile 3	0.08	-0.03	0.19	0.12
	(0.10)	(0.11)	(0.13)	(0.12)
Quartile 4	-0.10	0.07	0.10	0.04
	(0.11)	(0.12)	(0.17)	(0.13)

Table 5: Results by Baseline Test Scores - Urban Sample

Dependent Variables:	Science & Environment			
	Human Body	Environment	Physical World	Overall
	(1)	(2)	(3)	(4)
Panel C: Quartile Effects - Urban Boys				
Quartile 1	-0.11	0.19*	0.13	0.09
	(0.13)	(0.11)	(0.14)	(0.13)
Quartile 2	0.07	0.31***	0.15	0.21*
	(0.13)	(0.10)	(0.12)	(0.11)
Quartile 3	0.08	-0.00	0.38**	0.22
	(0.14)	(0.13)	(0.16)	(0.15)
Quartile 4	0.10	-0.00	0.36***	0.22*
	(0.08)	(0.12)	(0.13)	(0.12)
Observations	1718	1718	1718	1718

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Test scores expressed in standard deviations with respect to the control group.

Overall we observe a very similar pattern with respect to the entire sample. Column three of Panel A shows increasing impacts along the baseline distribution for the Physical World module. These impacts are again significant for the top half performers. Panel B shows null impacts for girls; while Panel C documents positive impacts for boys in the top half of baseline scores with respect to the Physical World module and the overall test score. Clearly, the results found for the entire sample are driven by the results achieved within urban schools.

4.6. Effects on other subjects

The new teaching method was not aimed at improving test scores in other academic subjects (mathematics and reading comprehension). However, this possible effect was incorporated into the pilot through baseline and end line application of exams in mathematics and reading comprehension. Estimated effects, however, suggest no statistically significant effects on these subjects.

The results indicate that at least in the short run the new pedagogical method does not impact learning in other academic subjects. However, it is still an open question whether the method could impact other learning areas over the long run.

4.7. The influence of the educational level of the parents

The heterogeneity of results could be partly due to differences in the characteristics of the household (Freeman et al, 2010). One key aspect of household characteristics is the education level of the household head. Table 6 shows the impacts differentiated by the level of education, that is, without secondary level of education or with secondary or higher level of education of the household head.

Table 6: Effects and the Educational Level of the Student's Household Head

Dependent Variables:	Endline Standardised Scores	
	Comparison Group Mean	Endline Score
	(1)	(2)
Panel A: Effects - Human Body		
HH head without	-0.20	0.13
Secondary		(0.14)
HH head with	0.23	-0.04
Secondary or higher		(0.11)
Observations		1109
Panel B: Effects - Environment		
HH head without	-0.11	0.08
Secondary		(0.18)
HH head with	0.24	0.08
Secondary or higher		(0.13)
Observations		1109
Panel C: Effects - Physical World		
HH head without	-0.14	0.30**
Secondary		(0.14)
HH head with	0.21	0.29*
Secondary or higher		(0.15)
Observations		1109
Panel D: Effects - Overall Science & Environment		
HH head without	-0.19	0.23
Secondary		(0.17)
HH head with	0.28	0.16
Secondary or higher		(0.15)
Observations		1109

Table 6: Effects and the Educational Level of the Student's Household Head

Dependent Variables:	Endline Standardised Scores	
	Comparison Group Mean	Endline Score
	(1)	(2)
Panel E: Effects - Math		
HH head without	-0.12	0.10
Secondary		(0.14)
HH head with	0.17	0.22
Secondary or higher		(0.14)
Observations		1115
Panel F: Effects - Reading		
HH head without	0.02	0.03
Secondary		(0.15)
HH head with	0.18	0.15
Secondary or higher		(0.11)
Observations		1107

Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Test scores expressed in standard deviations with respect to the control group.

The first column of the table shows the average standardised test scores differentiated by the educational level of the household head regarding the control group at end line. The standardised test scores have an average value of zero and a standard deviation of one. For virtually all subjects tested the average score is negative for students whose household head has lower than a secondary education, that is, students below the mean of the overall distribution. In contrast, students whose household head possess secondary or higher educational levels have positive average scores (above the overall mean). Thus students with more educated parents have higher test scores.

In column two of the table, the impacts of the program differentiated by the level of education of the household head are shown. A significant positive impact is found only in the Physical World module (panel C). However, the sizes of the impact are similar, 0.3 and 0.29 of a standard deviation, for household heads without and with secondary education respectively. So although there is a marked difference in test scores for

students according to the education level of the household head, the positive absolute impact of the program is independent of the different educational levels of the household head.

4.8. Perceptions of the teachers

The pilot also attempted to determine the impact of the intervention on the perceptions of the teachers regarding their own work and of their students. Potential answers to each question in the survey were in a scale ranging from one to five, where one denoted “very much in agreement” and five denoted “very much in disagreement” This scale was standardised between the interval [0; 1] such that if the specific perception had a negative connotation then a value of zero denoted “very much in agreement” while a value of unity denoted “very much in disagreement”. If the perception was on a positive characteristic then the value zero denoted “very much in disagreement” and the value of unity denoted “very much in agreement”.

The different areas of teacher perceptions covered in the questionnaire were: work in the school; knowledge of science and the environment; the importance of science and the environment; motivation in teaching science and the environment; and the process of teaching. Generally there was no impact of the intervention on teachers’ perceptions regarding their work. However, there was one significant exception. There was a negative impact on teachers’ perception that it is necessary first to teach theory and only afterwards practice. While the convenience of the timing between theory and practice is arguable, this question was intended to capture whether teachers have achieved what the methodology was looking for: to shift away from in class explanations to practical

examples using the LEGO kits. Therefore this result shows that the teachers in the treatment group at least in rhetoric successfully managed to shift away from the traditional focus on science history and the memorization of concepts.

In addition, the questionnaire was directed at the perceptions teachers had of their students. In particular, it evaluated the teachers' perceptions on students in the following areas: general motivation, working in teams, domestic relations (understood as teacher's perceptions regarding the support and commitment of student's families in the learning process), intellectual capacity and performance in science and the environment. The results are given in Table 7.

Table 7: Perceptions of the teachers regarding pupils

	Endline Comparison Group Mean	Program Impact	Number of Observations
	(1)	(2)	(3)
Panel A: Dependent Variables - General Motivation			
Have good behavior	0.66	0.01 (0.05)	124
Are very interested in learning	0.77	-0.01 (0.04)	125
Have very good attendance rate	0.77	-0.01 (0.05)	124
Average over family of outcomes (in standard deviations)		**-0.02 (0.12)**	**136**
Panel B: Dependent Variables - Teamwork			
Know how to work in team	0.73	0.04 (0.05)	124
Know to listen and speak in group	0.70	-0.03 (0.04)	124
Are autonomous	0.64	-0.05 (0.05)	124
Average over family of outcomes (in standard deviations)		**0.07 (0.08)**	**136**
Panel C: Dependent Variables - Domestic Relations			
Have problems at home that affect learning	0.49	0.02 (0.06)	124
Have the support of their families	0.41	0.07 (0.05)	123
Average over family of outcomes (in standard deviations)		**0.08 (0.14)**	**136**

Table 7: Perceptions of the teachers regarding pupils

	Endline Comparison Group Mean	Program Impact	Number of Observations
	(1)	(2)	(3)
Panel D: Dependent Variables - Intellectual Capacity			
Have good oral and written expression capacity	0.58	0.04 (0.04)	125
Have concentration problems that affect learning	0.42	-0.03 (0.05)	124
Have the capacity to learn any concept	0.60	0.02 (0.04)	123
Know how to reason and never study by memory	0.57	-0.03 (0.04)	124
Average over family of outcomes (in standard deviations)		**0.06** **(0.11)**	**136**
Panel E: Dependent Variables - Performance in Science & Environment			
Are very interested in learning science & environment	0.72	0.08* (0.04)	125
Have good performance in science & environment	0.66	0.06** (0.03)	124
Average over family of outcomes (in standard deviations)		**0.34**** **(0.14)**	**136**

Column (1) reports the average of the outcome within the control group at endline. Column (2) reports coefficients from one regression where the indicator for treatment school enter as RHS variable. Column (3) reports the number of observations in each regression. Estimated standard errors clustered at the school level are in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

The only significant effects were on teachers' perceptions of students' interest and performance in science and the environment (panel E). Teachers in the treated schools relative to those in non-treated schools perceived that students were interested in learning science and environment and had good performance in this subject. The impact, measured by standard deviations of the index, was 0.34 units. Therefore, it appears that teachers perceive their students more interested and with better performance in the subject. To the extent that teacher's perceptions are correlated with student's performance, we interpret these findings as corroborating evidence regarding the positive effects on standardised tests scores.

5) Conclusions

As Peru becomes closer to universal primary education the country is shifting attention from increasing coverage towards improving the quality of education. The quality of education in Peru is a major problem as evidenced by the country's low performance on national, regional and international standardised tests. The need for evidence of what improves the quality of education is a crucial policy issue.

In this paper we present the findings of an experimental evaluation of a new student centred pedagogical approach in the teaching of Science and Environment at the third grade level in Peru. The study draws on data from a rigorously designed intervention as well as from surveys of principals, teachers, students and parents. Despite complications resulting in only a partial implementation of the pilot we find positive and significant improvements in test scores of students taught with the new method. These effects were concentrated in the geographical areas were the program was more intensively implemented (urban schools) and were stronger for students that were ex-ante better off. In addition, it appears that as boys monopolized the didactic materials, they were the ones that were differentially benefited by the program.

A challenge for an eventual scale-up of the program consists in ensuring that all students benefit from the new teaching methods, particularly girls and students in rural schools. This is a critical policy design issue in a country with an inequitable distribution in the quality of education and learning outcomes.

[i] For instance, between 1998 and 2009 enrolment in pre-school education increased from 53.4% to 66.3%; enrolment in primary education increased from 90.6% to 94.4%; enrolment in secondary education increased from 59% to 76.5% (source: Peruvian Ministry of Education)

[ii] Evaluación Censal de Estudiantes, UMCE; MINEDU 2010.

[iii] Minimum wage in Peru during the pilot was S/.550 per month.

[iv] Notice that we express the function in a linear specification for ease of exposition. However, the unknown production function may take several types of non linear forms.

[v] Notice, however, that there is an extensive non-experimental literature using simulation techniques with the objective of estimating education production functions and the relative incidence of different cognitive and non cognitive inputs (Cunha and Heckman, 2007)

[vi] Notice, however, that if the choosing of two sections within schools with more three sections or more would not have been really random and, furthermore, this would have not been balanced between treated and control schools then our research design would be biased. However, as we show in Table 1, schools are balanced between treated and control groups regardless of the number of sections that the school has.

[vii] Validation of the design of the exams for Science and the Environment, Reading Comprehension and Mathematics, had been tested previously in two schools in Lima that are not part of the sample. Baseline tests had been applied to third grade pupils in the beginning of the school year and the end line at the end of the school year. While the pilot program did not include specific activities to improve reading and mathematics skills, it was considered desirable to also evaluate these areas. These exams covered the learning expectations on each subject at each point in time. These were standard exams that covered the areas that were supposed to be learned according to the national curriculum. Recall that our methodology only changed the way of teaching but learning targets were the same under the traditional and the new experimental methodology being tested.

[viii] Notice that all test scores are expressed in standard deviations with respect to the control group. That is, the control group test scores have mean zero and standard deviation of unity.

[ix] At baseline, the total number of students was 2,802, however from this number were excluded pupils pertaining to the inclusion program (a program that integrates students with special education needs in regular classrooms), those that moved to the fourth grade during the school year, and those that did not finish all the tests; leaving 2,771 effective observations. At the end line, the total number of students tested were 2,663, excluding pupils of the inclusion program, those that had moved to the fourth grade, and those that did not finish all the tests the figure falls to 2,552. Of the latter group only 2,373 pupils were evaluated at baseline.

[x] The exam results were standardised such that the control group had a zero mean and a standard deviation of unity. Therefore, differences between treated and control groups are expressed in standard deviations with respect to the control group.

[xi] Again, exam results were standardised such that the control group had a zero mean and a standard deviation of unity. Therefore, differences between treated and control groups are expressed in standard deviations with respect to the control group.

[xii] We also did such exercise for rural schools. However, no impacts at all were found.

References

Banerjee, A., Cole, S., Duflo, E. and Linden, L. (2007) Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics*, 122(3), pp. 1235-1264.

Barrera-Osorio, F. and Linden, L. (2009) The use and misuse of computers in education : evidence from a randomized experiment in Colombia. *World Bank Policy Research Working Paper 4836.*

Benavides, M. and Mena, M. (2010) Informe de progreso educativo, Perú 2010. GRADE, Peru.

Bloom, H. and Michalopoulos, C. (2010) When is the story in the sub groups: strategies for interpreting and reporting effects for sub groups. *MDRC Working Paper 551*.

Bloom, H., Bos, J. and Lee, S. (1999) Using cluster random assignment to measure program impacts: statistical implications for the evaluation of education programs. *Evaluation Review*, 23 (4), pp. 445-469.

Bonwell, C. and Eison, J. (1991) Active Learning; Creating Excitement in the Classroom. *AEHE-ERIC Higher Education Report No. 1.* Washington DC.

Carrillo, P., Onofa, M. and Ponce, J. (2010) Information Technology and Student's Achievement: Evidence from a Randomized Experiment in Ecuador. Unpublished document. Washington, DC: Inter-American Development Bank.

Colburn, A. (2000) An inquiry primer. *Science Scope*, 23 (6), pp. 42-44.

Cristia, J., Cueto, S., Ibarraran, P., Severin, E., and Santiago, A. (2012) Technology and Child Development: Evidence from the One Laptop per Child Program. *IDB Working Paper No 273.*

Cunha, F., and Heckman, J. (2007) The Technology of Skill Formation. *American Economic Review*, 97(2), pp 31-47.

Duflo, E., Dupas, P., and Kremer, M. (2011) Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5), pp. 1739–1774.

Freeman R., Machin, S. and Viarengo, M. (2010) Variation in educational outcomes and policies across countries and schools within countries. *CEE Discussion Papers 0117*.

Furtak, E., Seidel, T., Iverson, H., and Briggs, D. (2012) Experimental and Quasi-Experimental Studies of Inquiry-Based Science Teaching: A Meta-Analysis. *Review of Educational Research*, 82 (3), pp. 300-329.

Glewwe, P. and Kremer, M. (2006) Schools, teachers, and education outcomes in developing countries, in: Hanushek E. and Welch, F. (eds.) *Handbook of the economics of education* (Elsevier), 2(2).

Glewwe, P., Hanushek, E., Humpage, S., and Ravina, R. (2011) School resources and educational outcomes in developing countries: A Review of the Literature from 1990 to 2010. *Unpublished manuscript, University of Minnesota*.

Healy, J. (1990) Endangered minds: why our children don't think and what we can do about it. (New York: Simon and Schuster).

Iturrizaga, I. (2000) Study of the educational impact of the LEGO Dacta Materials-INFOESCUELA-MED. Peruvian Ministry of Education, pp. 1-39.

Hussain S., Lindh, J. and Shukur, G. (2006) The effect of Lego training on student's school performance in mathematics problem solving ability and attitude: Swedish data. *Educational Technology and Society*, 9(3), pp. 182-194.

Lowery, L. (1998) The biological basis of thinking and learning. University of California, Berkeley.

Malamud, O. and Pop-Eleches, C. (2011) Home Computer Use and the Development of Human Capital. *Quarterly Journal of Economics*, 126 (2), pp. 987-1027.

Ministerio de Educación. (2008) Diseño Curricular Nacional de Educación Básica Regular, Lima - Peru.

Noble M. (2001) The Educational Impact of Lego Dacta Materials. Sheffield Hallam University.

Organization for Economic Cooperation and Development (OECD) (2011) PISA 2009 Results: What Students Know and Can Do. Paris: OECD

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2008) First Report of the Second Regional Comparative Education Study. Santiago de Chile, Chile.