NBER WORKING PAPER SERIES

EFFECT OF INQUIRY AND PROBLEM BASED PEDAGOGY ON LEARNING: EVIDENCE FROM 10 FIELD EXPERIMENTS IN FOUR COUNTRIES

Rosangela Bando
Emma Näslund-Hadley
Paul Gertler

Working Paper 26280
http://www.nber.org/papers/w26280

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2019

The authors gratefully acknowledge the outstanding research assistance of Gabriel Englander and Harold Villalba. The authors thank the Inter-American Development Bank and the Government of Japan for funding the research. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the Inter-American Development Bank, its Board of Directors, the countries they represent, or the National Bureau of Economic Research. The authors have no conflicts of interests or financial or material interests in the results.

Effect of Inquiry and Problem Based Pedagogy on Learning: Evidence from 10 Field Experiments in Four Countries

Rosangela Bando, Emma Näslund-Hadley, and Paul Gertler

NBER Working Paper No. 26280
September 2019
JEL No. I21,I24,I25

ABSTRACT

This paper uses data from 10 at-scale field experiments in four countries to estimate the effect of inquiry-and problem-based pedagogy (IPP) on students' mathematics and science test scores. IPP creates active problem-solving opportunities in settings that provide meaning to the child. Students learn by collaboratively solving real-life problems, developing explanations, and communicating ideas. Using individual-level data on 17,006 students, the analysis finds that after seven months IPP increased mathematics and science scores by 0.18 and 0.14 standard deviations, respectively, and by 0.39 and 0.23 standard deviations, respectively, after four years. We also identify important gender learning gaps with boys benefiting substantially more than girls. Our approach not only provides strong causal evidence, but also high external validity. These 10 experiments in four countries allow us to examine the effects of IPP across a wide set of geographic, socioeconomic, teacher background, and age/grade contexts (i.e., preschool and third and fourth grades). The results prove to be robust across these different contexts. The 10 RCTs were registered in the American Economic Association Registry for randomized control trials. See the supplementary materials for trial numbers.

Rosangela Bando
Inter-American Development Bank
1300 New York Avenue N.W.
Washington, DC 20577
rosangelab@iadb.org

Emma Näslund-Hadley
Inter American Development Bank
1300 New York Avenue N.W.
Washington, D.C. 20577
EMMAN@iadb.org

Paul Gertler
Haas School of Business
University of California, Berkeley
Berkeley, CA 94720
and NBER
gertler@haas.berkeley.edu

The 10 RCTs were registered in the American Economic Association Registry for randomized control trials. See the supplementary materials for trial numbers.

1. Introduction

The education literature has long emphasized that students learn better when they play an active role in the learning process through do-able tasks with social interaction (*1, 2, 3, 4*). Meta-analyses confirm that traditional lecturing with passive listening is not conducive to critical thinking, fostering interest, or changing attitudes (*5, 6*). Rather, learning through activities, group work, and interactive class conversations is strongly associated with greater learning (*7*).

One such active leaning approach is inquiry- and problem-based pedagogy (IPP) (*8*). IPP creates active problem-solving opportunities in settings that provide meaning to the child. Students learn by collaboratively solving authentic, real-life problems, developing explanations, and communicating ideas (*9*). They are taught to search for information from different sources, both text-based resources and by gathering their own data, and to develop problem-solving skills by collaboratively engaging in investigations. This approach helps solidify concepts through the child's exploration of research questions, production and collection of evidence, construction of theories based on evidence, and development of explanations.

This paper uses student-level data from 10 randomized field experiments in four Latin American countries (Argentina, Belize, Paraguay, and Peru) to estimate the effect of IPP compared to traditional pedagogy on preschool and primary school student learning in mathematics and science (*10*). We estimate both short-run and longer-run effects considering that learning begets learning, i.e., dynamic complementarities (*11*). The results show that the longer-run impact is significantly larger, increasing the cost-effectiveness of IPP. Finally, the analysis finds that boys benefit more than girls from IPP and that the gender gap grows over time.

Our approach not only provides strong causal evidence, but also high external validity. A challenge when evaluating specific programs is the applicability of the evidence to other contexts (*12, 13*, *14*). These 10 experiments in four countries allow us to examine the effects of IPP across a wide set of geographic, socioeconomic, teacher background, and age/grade contexts (i.e., preschool and third and fourth grades).

2. Inquiry- and Problem-based Pedagogy

The difference between IPP and a traditional lesson is illustrated by a unit on the skeletal system in the fourth grade in Argentina (*15*). In traditional classrooms, students copy facts about bone tissues and the names of the 206 bones of the human skeleton that teachers have written on the blackboard into notebooks. They are then tested based on the lectures and material that they have read in textbooks. In IPP classrooms, teachers pose research questions and guide students through the formulation and testing of hypotheses to explore the questions. One research question might be: What do bones help people do? Students then research facts about bones from texts and other sources from which they devise hypotheses. One such hypothesis is that calcium strengthens bones. Students might then soak chicken bones in vinegar for different lengths of time to extract different amounts of calcium, concluding that the more calcium a bone loses, the more it will bend.

In mathematics, the contrast between IPP lessons and traditional lessons is equally stark. For example, consider a lesson on ratios in the sixth grade in Belize (*16*). In a traditional classroom, the lesson begins with a lecture that covers the definition of a ratio and how to solve simple mathematics problems involving ratios. The students then spend the rest of the class solving similar problems and are tested on their ability to solve ratio problems. In an IPP classroom, the teacher first uses examples to convey the concept (e.g., the ratio of students with long-sleeve shirts to those with short-sleeve shirts). Students then work in pairs to come up with definitions. The teacher provides them with a series of exercises to explore the use of ratios in everyday life. For example, pairs of students might be asked to investigate how many Cuisenaire rods of different colors are needed to measure the length of their desks and the relationships between the numbers of rods of different colors (*17*). The small group exploration is followed by a teacher-led class discussion. The lesson ends with students revising their definitions of a ratio and a class conversation guided by the teacher to arrive at a joint definition and properties.

Teachers play critical roles in IPP. When done well, IPP includes elements of explicit instruction and scaffolding (*18,19*). Teachers facilitate learning by guiding students through a series of steps and explicitly relating learning to students' prior knowledge and experiences (*18*). Teachers guide learners through complex tasks with explicit instructions that are relevant to the problem at hand (*19*). They provide structure and scaffolding that help students not only carry

out specific activities, but also comprehend why they are doing those activities and how they are related to the set of core concepts that they are exploring (*1*).

3. The Interventions

This study encompasses 10 IPP randomized field experiments in four Latin America countries: Argentina, Belize, Paraguay, and Peru. The countries represent GDP per capita income levels that range from US$4,078 in Paraguay to US$12,440 in Argentina, and they range in population sizes from 366,954 in Belize to 43,847,430 in Argentina (*20*). Like many countries, these four nations face challenges with education quality, as illustrated by their national and international scores that show severe learning deficits compared to Organisation for Economic Co-operation and Development countries (*21*, *22*). Supplementary Table S1 provides details of each of the 10 IPP interventions.

All interventions shared three central elements of IPP: (1) instruction organized around core concepts that were developed over many lessons, (2) classes organized around inquiry and problem-solving opportunities, and (3) use of students' previous knowledge, structure, and scaffolding to help them carry out more complex activities and make sure that they have close guidance.

All programs were implemented at the class level, except for Peru 2014, where tutors were used for small groups of three to seven students. Each program trained teachers in IPP methods and lesson plans, provided didactic materials to enhance learning through hands-on activities, and provided ongoing supervision. All programs included detailed lesson plans, a minimum of 20 hours of teacher professional development, and continuous in-school teacher support.

4. Experimental Designs

Although the details of each study differ, all studies employed a cluster (school-level) randomized design, except for Peru 2014. Peru 2014 randomized students at the individual level. Study schools in Argentina and Peru were randomly selected from the respective country-year universe of schools with students enrolled in the grade of interest. In Paraguay and Belize, study schools were selected from the universe of eligible schools that had students in the grade of interest and that additionally volunteered to participate. Schools were compliant with treatment assignment in all cases except for one control school in Paraguay 2011 where teachers received training. For this case, we present intention to treat estimates. Except for Peru 2014, all students

in the target grades in the study schools participated in the study. Peru 2014 instead enrolled students who performed in the bottom half of the test score distribution. Supplementary Table S1 provides details of each of the 10 IPP interventions, and Supplementary Table S2 provides the details of each experimental design, including sample frame, sample size in terms of number of schools and number of students, stratifications for random assignment, and timing of data collection.

All studies except for Belize 2015 collected panel data at the student level with one survey before treatment and another after treatment. In all studies the same group was surveyed before and after the intervention, except for Belize 2015, where baseline and follow-up surveys were administered to different cohorts. The length of exposure to inquiry- and problem-based pedagogy (IPP) was seven months in all cases.

The key outcome of interest is students' standardized test results. Each test was designed to measure the ability of students to understand and apply key mathematical and scientific concepts. Tests were adapted for each grade level and administered by an external evaluator, rather than by the local teachers. Surveys of parents provided additional information about the student and family. Teacher and school-level information was merged into the student-level data base. Supplementary Table S3 provides the definition for each variable used in the analysis.

5. Estimation

We estimate the following regression specification for each country-year subject intervention:

$$y_{ist} = \mu_s + \beta T_{is} + \gamma y_{ist-1} + \varepsilon_{is}, \quad (1)$$

where y_{ist} denotes the score for student *i* in strata *s* at time *t*, μ_s is a strata fixed effect, and ε_{is} is an error term. The variable T_{is} equals 1 if the student receives treatment and 0 otherwise. β represents the average difference in student scores between treatment and control units in the year in which IPP was implemented. For inference, we cluster errors at the school level (*23*).

An importation notion of learning is that how much a child learns in a school year depends on how much he or she knows upon entering that year, i.e., school readiness. These dynamics are built into equation (1), where current test scores are also a function of lagged scores, γy_{ist-1} (*24*). A key implication is that an intervention that improves learning today will

improve learning in future periods. Hence, just evaluating the contemporaneous impact of IPP underestimates the full impact.

Specifically, the impact of IPP in year *t* on learning in year *t* is β, the impact of IPP in year *t* on learning in year *t*+1 is $\gamma\beta$, the impact of IPP in year *t* on learning in year *t*+2 is $\gamma^2\beta$, etc. We can then obtain the estimated full impact by summing up the years since intervention. The impact of one year of IIP after four years is $\beta(1 + \gamma + \gamma^2 + \gamma^3)$. We use the delta method to compute standard errors (SE). This assumes that γ does not change across grades, and there is some evidence to support this assumption in that we cannot reject pooling across the samples that are representative of grades kindergarten through fourth grade.

We can also estimate the impact of multiple years of IPP. In this paper we estimate what would happen if primary schools were to completely shift to IPP for grades one through four. This could be estimated by $\beta(4 + 3\gamma + 2\gamma^2 + \gamma^3)$. This assumes that both β and γ do not change across grades, and there is some evidence to support this assumption in that we cannot reject pooling across the samples that are representative of grades kindergarten through fourth grade.

6. Results

The supplemental material provides descriptions of each of the randomized experimental designs and is summarized in Supplementary Table S2. The variables used in the analysis are described in Supplementary Table S3, with baseline balance and attrition assessed in Supplementary Tables S4 and S5, respectively.

Baseline Balance and Sample Attrition

Descriptive statistics at baseline prior to the interventions, and p-values for tests of the hypotheses that the means of the treatment group are equal to those of the control group, show that the treatment and control groups are well balanced for all the study samples (Supplementary Table S4). Mean mathematics and science test scores in the treatment group are not statistically different from the control group for all countries and years. Similarly, there are no differences for student age, whether bilingual, family assets, teacher's age, and gender. However, there are significant gender imbalances in the Belize 2015 and Argentina 2009 science experiments, and in class size in the Belize 2015 and Argentina 2009 mathematics and science experiments.

The attrition rates by treatment and control groups, for each country (except for Belize, for which we do not have a panel of students) show little evidence of selective attrition bias (Supplementary Table S5). Student attrition over the seven-month period ranges from 3 percent in Paraguay 2011 to 17 percent in Argentina 2009. There is no differential attrition between treatment and control groups for all study samples except for Argentina 2009, where there was 4 percentage points more attrition in the control group than in the treatment group. Despite this, there appears to be no differences in the means of baseline test scores between treatment and control groups for the evaluation sample, i.e., the sample was found at endline (Supplementary Table S6). Overall, we can reject only five of the 64 tests of the equality of treatment and control means at the 0.10 significance level.

Pooling Tests

The estimation results of equation (1) for each of the 10 study samples are presented in Supplementary Table S6. We also estimate equation (1) with a common β and γ across all samples, but allowing the strata dummies to vary by country and year (Table 1). We cannot reject that the coefficients are not different across the samples for mathematics, science, or both using *F*-tests. *P*-values for the *F*-tests are presented in row 3 of Table 1.

We also take a meta-analysis approach to construct an average of the individual country-year estimates weighted by the inverse of the variance of the estimate (*25*). We test for cross-study heterogeneity using an I^2 statistic, which measures the percentage of variation attributable to heterogeneity across studies (*26*). I^2 takes values between 0 and 100 percent, with 100 percent indicating high heterogeneity across studies (*27*). The I^2 for studies within mathematics, science, and overall is 0 percent, implying that we cannot reject the hypotheses that the estimated coefficients are equal across all study samples for mathematics ($p = 0.828$) and for science ($p = 0.728$).

Short-run effects

The short run (seven-month) impact of IPP shows meaningful positive and statistically significant effects on both mathematics and science test scores (β rows in Supplementary Table S6 and Table 2) (*28*). The short-run impact on mathematics scores is 0.18 standard deviations (SD) overall and ranges from 0.13 SD in Argentina 2009 to 0.20 SD in Paraguay 2011. The

impact on science scores is 0.14 SD and ranges from 0.08 SD in Argentina 2009 to 0.29 SD in Belize 2015. Figure 1 depicts the pooled and country-year estimates.

Longer-run Effects

Table 1 also reports estimates of γ. The results show that dynamic complementarities are important, as 1 standard deviation of knowledge entering the grade translates into an additional 0.58 SD of learning in mathematics and an additional 0.39 SD in science. Taking these dynamics into account, we estimate that after four years, the impact is 0.39 SD in mathematics and 0.23 SD in science (Table 2). Supplementary Table S7 provides these results for each of the 10 samples. Accounting for dynamics more than doubles the estimated impact on mathematics learning and increases it by over 60 percent in science. In addition, the accumulated learning impact of four years of IPP is 1.21 SD mathematics and 0.79 SD in science.

Gender Differences

Separate estimation by gender reveals that boys benefit significantly more from IPP than girls (Table 3). The instantaneous treatment coefficient β *i*s 0.22 SD for boys versus 0.15 SD for girls in mathematics and 0.18 SD for boys versus 0.10 SD for girls in science. Moreover, the effect is statistically significantly different. However, the effect of lagged test scores, γ, is the same for boys and girls for both mathematics and science.

Gender gaps in short-run impacts do translate into substantially different treatment effects in the long run (Table 4). The male-female gap in terms of impact from one year of treatment is even larger after four years, growing from 0.07 SD in the first year to 0.17 SD in the fourth year for mathematics and from 0.08 SD to 0.15 SD in science over the same period. In addition, the gender gap from four years of treatment is even larger: in mathematics the gender gap would be 0.49 SD and in science 0.50 SD.

Cost-effectiveness

Finally, we provide estimates of cost-effectiveness using administrative data for each program to estimate incremental costs. We use the Consumer Price Indices for All Urban Consumers to normalize the costs to March 2017. We include teacher training, didactic materials, and supervision costs. Training and material costs are depreciated over a three-year period using straight-line depreciation.

We calculate the cost of a 0.10 SD increase (Table 2). We find that the cost of increasing test scores by 0.10 SD after one-year is US$18.12 per student in mathematics and US$17.89 in science. However, when we estimate the four-year impact, the cost of a 0.10 SD increase in scores falls to US$8.37 in mathematics and US$10.89 in science (*29*). Supplementary Table S7 provides these results for each of the 10 samples.

7. Discussion

This paper has analyzed data from 10 field experiments in four countries to assess if teacher training designed to change pedagogical practices from teacher-centered lecturing with passive listening to student-centered IPP learning processes improved student test scores. Our results strongly support the conclusion that implementing IPP enhances student learning in mathematics and science.

After one school year of IPP, mathematics test scores increased by 0.18 SD and science scores increased by 0.14 SD. Accounting for dynamic complementarities, the estimated effects of one year of IPP rise to 0.39 SD in mathematics and 0.23 SD in science after four years. The effects of IPP on learning are likely to be lower-bound estimates of the true effect. This was the first time any of the teachers implemented IPP, and they would likely improve their IPP teaching skills over time.

IPP benefited boys more than girls by 0.07 SD in mathematics and 0.08 SD in science after one year. After four years, the male-female gap increased to 0.17 SD in mathematics and 0.15 SD in science.

A major finding is that the effect sizes were not different in order of magnitude or statistical significance across the 10 experimental settings, suggesting a greater degree of external validity than most studies. This is important because programs varied in terms of setting, intensity, complementary learning materials, and teacher support. These results were present across two subject areas (mathematics and science), three grade levels (preschool and third and fourth grades), and four countries and educational systems. Teachers had different backgrounds. The 2014 Science Program in Peru showed effects when IPP was implemented as a tutoring program outside of the classroom. Further, the programs targeted students in different sociocultural conditions.

The cost of scaling the IPP approach is low and decreases once we account for dynamic complementarities. We estimate that the costs of increasing test scores by 0.10 SD decrease between years one and four from US$18.12 to US$8.37 per student in mathematics and from US$17.89 to US$10.89 per student in science.

Our results are broadly consistent with the previous IPP literature. Qualitative assessments of the programs we studied found that classes were more interactive, and students were more involved in academic activities in treatment schools than their peers in control schools (*30, 31, 32*). Our findings are also in line with the education literature that suggests that some degree of inquiry-based classroom practices enhances learning (*2*) and that guided inquiry is more effective than minimally guided instructional approaches (*3*).

Finally, our results are consistent with studies of individualized instruction more generally as a pedagogical approach. A teacher training program that aimed to promote a student-centered pedagogical approach led to an increase of 0.25 SD in test scores after one year among fourth graders in secular schools (but had no effect in religious schools) in Jerusalem (*33*). Substituting two hours of class lecture per school day with individualized tutoring led to improvements of 0.14 SD after one academic year among first graders in India (*34*). Tracking students by ability increased learning by 0.16 SD after 18 months among first grade students in Kenya (*35, 36*).

References and Notes:

1. L.S. Vygotsky. *Mind in Society: The Development of Higher Mental Processes* (Harvard University Press, Cambridge, MA, 1978).
2. L.F. Lowery. *The Biological Basis of Thinking and Learning* (University of California, Berkeley, 1998).
3. E. Furtak, T. Seidel, H. Iverson, D. Briggs. Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research* **82**(3), 300-329 (2012).
4. More generally, there is broad consensus that a teacher's pedagogical style and the quality of teacher-student interactions are key inputs into student learning. See (*37*) and (*38*).
5. D. Bligh. *What's the Use of Lectures?* (Jossey-Bass, San Francisco, 2000).
6. E. Näslund-Hadley, A. Loera Valera. K.A. Hepworth. What goes on inside Latin American math and science classrooms: A video study of teaching practices. *Global Education Review* **1**(3), 110-128 (2014).
7. S. Freeman, S.L. Eddy, M. McDonough, M.K. Smith, N. Okoroafor, H. Jordt, M.P. Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences* **111**(23), 8410-8415 (2014).
8. While inquiry- and problem-based learning have different origins, they are similar in practice. Inquiry learning has its roots in scientific research (*39*), and problem-based learning has its roots in medical education (*40, 41*).
9. C.E. Hmelo-Silver. Problem-based learning: What and how do students learn? *Educational Psychology Review* **16**, 235-266 (2004).
10. The program implemented in Peru in 2010, investigated in (*42*), found that IPP led to a 0.18 standard deviation increase in scores.
11. F. Cunha, J. Heckman**.** The technology of skill formation. *American Economic Review* **97**(2), 31-47 (2007).
12. C.F. Manski. *Public Policy in an Uncertain World* (Harvard University Press, Cambridge, MA, 2013).
13. S. Athey, Susan, G.W. Imbens. 2017. "The econometrics of randomized experiments" in *Handbook of Field Experiments,* E. Duflo, A. Banerjee, Eds. (North Holland, 2017), vol.1, ed. 1.
14. An active research area is the identification of relevant dimensions and relationships to extrapolate lessons learned. For example, (*43, 44, 45*) study extrapolation error.
15. Inter-American Development Bank. AR-T1047: Improvement of natural science and mathematics education (2018a). https://www.iadb.org/en/project/AR-T1047 (accessed March 14, 2018).
16. Inter-American Development Bank. BL-L1018: Education quality improvement (2018b). https://www.iadb.org/en/project/BL-L101 T1047 (accessed March 14, 2018).

17. Cuisenaire rods are colored wood pieces of different lengths used to visualize mathematics concepts.

18. C.E. Hmelo-Silver, R.G. Duncan, C.A. Chinn. Scaffolding and achievement in problem-based and inquiry-based learning: A response to Kirschner, Sweller, and Clark. *Educational Psychologist* **43**(2), 99-107 (2007).

19. D.C. Edelson. Learning-for-use: A framework for integrating content and process learning in the design of inquiry activities. *Journal of Research in Science Teaching* **38**, 355–385 (2001).

20. World Bank. World Development Indicators (2018). https://data.worldbank.org (accessed March 4, 2018).

21. M.S. Bos, A. Elías, E. Vegas, P. Zoido. "PISA Latin America and the Caribbean: How much did the region improve?" (Brief 2, Inter-American Development Bank, 2016).

22. The Programme for International Student Assessment (PISA) is an international survey conducted by the Organisation for Economic Co-operation and Development that evaluates problem-solving and cognition among 15-year-old students worldwide. http://www.oecd.org/pisa/.

23. For Argentina, we estimate confidence intervals with a bootstrap approach resampling from schools rather than students following (*46*).

24. Another reason to include lagged individual test scores is to reduce residual variance (*47*).

25. J.A. Sterne (editor). 2009. *Meta-Analysis in Stata: An Updated Collection from the Stata Journal* (Stata Press, College Station, TX, 2009).

26. J.P.T. Higgins, S.G. Thompson, J.J. Deeks, D.G. Altman. 2003. Measuring inconsistency in meta-analyses. *British Medical Journal* **327**, 557–560.

27. More specifically, I^2=100%*(*Q-df*)/*Q,* where *Q* is the across study variation of impacts, and *df*=*k*-1 denotes the degrees of freedom.

28. The point estimates are robust to the specific method used and estimates become more precise in models that add covariates and student fixed effects (Supplementary Table S7).

29. We exclude Peru 2014 from the overall cost-effectiveness analysis as it is a small group tutoring program and costly relative to the other interventions. Supplementary Table S8 lists cost-effectiveness estimates by country and subject.

30. UNESCO, Universidad Catolica. Programa de Mejora de la Enseñanza de las Ciencias Naturales y la Matemática. Volumes II, III and IV (2010).

31. N. Benson. **"**Tikichuela: Matemáticas en Mi Escuela documento anexo componente Cualitativo 2014" (Innovations for Poverty Action and the Inter-American Development Bank, Asunción, Paraguay, 2014).

32. Innovations for Poverty Action, Inter-American Development Bank (IDB), Ministry of Education, Peru. "Análisis cualitativo de la evaluación experimental del programa MIMATE y del programa de mejora de la Educación de Ciencias II" (Informe Final de Evaluación. IDB Consultancy Report (2014a).

33. J.D. Angrist, V. Lavy. 2001. Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics* **19**(2), 343-369 (2001).

34. A. Banerjee, S. Cole, E. Duflo, L. Linden. Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics* **122**(3), 1235-1264 (2007).

35. E. Duflo, P. Dupas, M. Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence form a randomized evaluation in Kenya. *American Economic Review* **101**, 1739-1774 (2011).

36. Tracking goes beyond tailored instruction because it involves peer effects.

37. R.J. Murnane, A.J. Ganimian. "Improving educational outcomes in developing countries: Lessons from rigorous evaluations" (NBER Working Paper No. 20284, National Bureau of Economic Research, Cambridge, MA, 2014).

38. M. Kremer, C. Brannen, R. Glennerster. The challenge of education and learning in the developing world. *Science* **340**, 297 (2013).

39. J. Dostál. *Inquiry-based Instruction: Concept, Essence, Importance and Contribution* (Palacký University, Olomouc, 2015).

40. H.S. Barrows, R.M. Tamblyn. *Problem-based Learning* (Springer Press, New York, 1980).

41. H.G. Schmidt. Problem-based learning: rationale and description. *Medical Education* **17**, 11-16 (1983).

42. D.W. Beuermann, E. Naslund-Hadley, I.J. Ruprah, J. Thompson. The pedagogy of science and environment: Experimental evidence from Peru. *The Journal of Development Studies* **49**(5), 719-736 (2013).

43. R. Dehejia, C. Pop-Eleches, C. Samii. From local to global: External validity in a fertility natural experiment. *Journal of Business and Economic Statistic*s (published online July 5, 2019; not yet published in print).

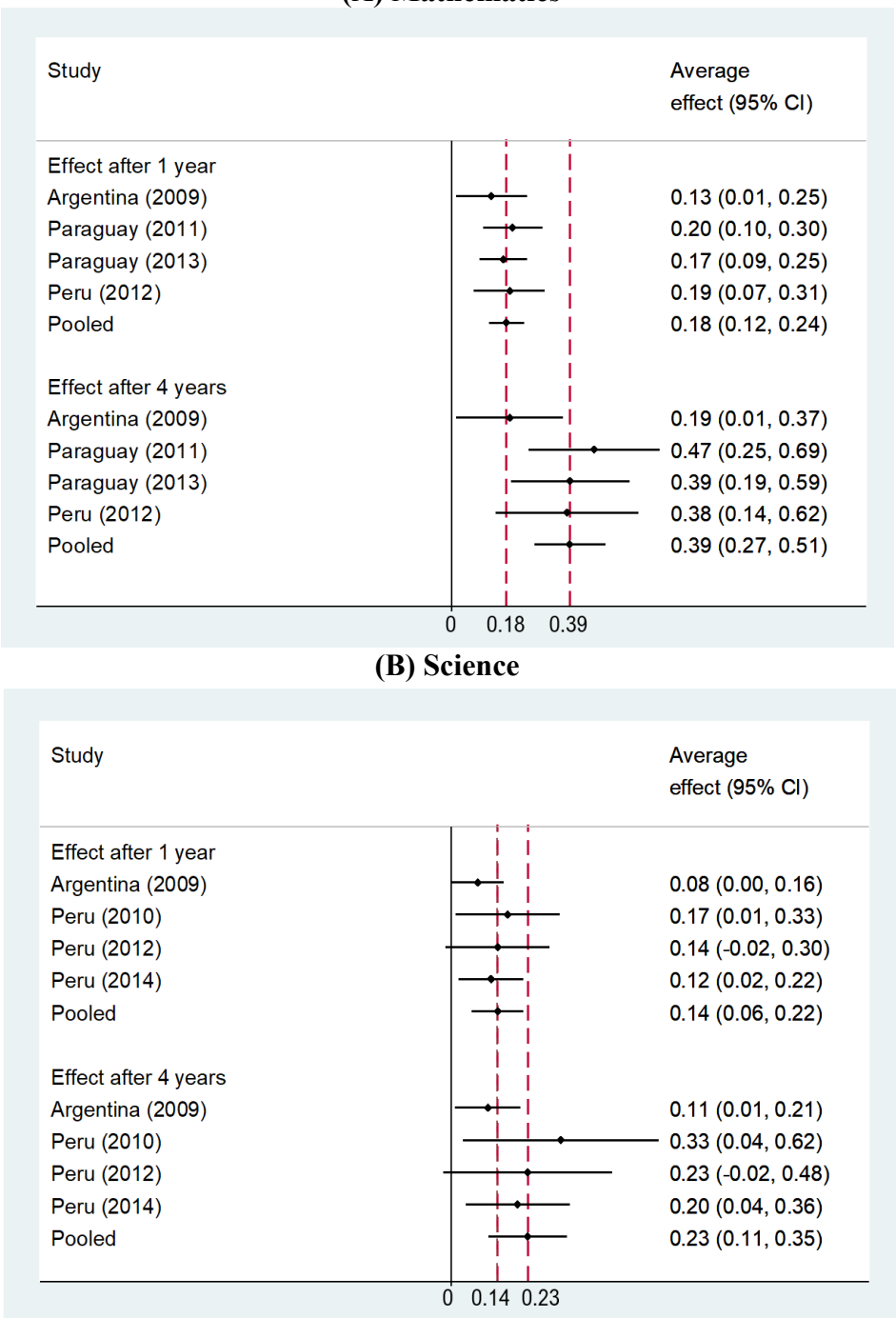
44. A. Hunt. Site selection bias in program evaluation. *Quarterly Journal of Economics* **130**(3), 1117-1165 (2015).

45. E. Vivalt. How much can we generalize from impact evaluation results? (Australian National University, 2014).

46. C.A. Cameron, D.L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources* **50**(2), 317-373 (2015).

47. W.G. Imbens, J.M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* **47**(1), 5-86 (2009).

Figure 1. The Impact of Inquiry- and Problem-based Pedagogy on Student Performance



Source: Prepared by the authors.

Note: This figure presents the pooled estimated impact and 95 percent confidence regions of inquiry- and problem-based pedagogy on mathematics and science tests in standard deviations for each/country year and by subject. CI: confidence interval.

Table 1: Pooled Estimates of Equation (1) by Subject and Combined

	Mathematics		Science		Combined	
	Pooled (1)	**Meta (2)**	**Pooled (3)**	**Meta (4)**	**Pooled (5)**	**Meta (6)**
β (Treatment)	0.18 (0.03) [0.000]	0.17 (0.05) [0.000]	0.14 (0.04) [0.000]	0.11 (0.06) [0.000]	0.16 (0.02) [0.000]	0.14 (0.04) [0.000]
γ (Lagged test score)	0.58 (0.01) [0.000]	0.58 (0.02) [0.000]	0.39 (0.02) [0.000]	0.33 (0.03) [0.000]	0.49 (0.01) [0.000]	0.50 (0.02) [0.000]
p-value for test of pooling	0.828		0.728		0.634	
Number of students	9,219		7,847		17,066	
Number of schools	659		300		959	

Note: Columns (1), (3), and (5) report the estimates of the model in equation (1) for the pooled samples. Columns (2), (4), and (6) report the meta-analysis estimates. Reported are the estimated coefficients, standard errors in parentheses, and p-value for the hypothesis that the coefficient equals zero in brackets. Standard errors and p-values are clustered at the school level. Belize 2015 is excluded from this analysis because it relies on cross-sectional data.

Table 2: Effect of One Year of IPP on Test Scores and Cost-Effectiveness by Time of Exposure

	Mathematics		Science	
Years Since Treatment (1)	Impact on Test Scores (2)	Cost per Student for 0.10 SD Increase in Test Scores (3)	Impact on Test Scores (4)	Cost per Student for 0.10 SD Increase in Test Scores (5)
1	0.18	$18.12	0.14	$17.89
	(0.03)		(0.04)	
2	0.29	$11.25	0.20	$12.52
	(0.04)		(0.05)	
3	0.35	$9.32	0.22	$11.38
	(0.05)		(0.06)	
4	0.39	$8.37	0.23	$10.89
	(0.06)		(0.06)	

Note: Columns (2) and (4) show the effects (and standard errors computed by the delta method) of one additional year of inquiry- and problem-based pedagogy (IPP) after the number of years listed in column (1). Columns (3) and (4) show the average yearly cost of IPP for a 0.10 standard deviation (SD) increase in test scores. Cost is the weighted average of the cost of the programs in Argentina, Paraguay, and Peru, excluding the tutoring program in Peru 2014.

Table 3: Overall Estimates of Equation (1) by Subject and Gender

	Mathematics		Science	
	Boys	**Girls**	**Boys**	**Girls**
β (Treatment)	0.22 (0.03)	0.15 (0.03)	0.18 (0.04)	0.10 (0.05)
γ (Lagged test score)	0.59 (0.01)	0.57 (0.02)	0.39 (0.02)	0.39 (0.02)
p-value boys = girls	0.000		0.001	

Note: Each column represents the estimated parameters for equation (1) within each gender group and subject. Standard errors listed in parentheses. The standard errors are clustered by school.

Table 4: Effect of One-Year of IPP on Test Scores by Gender

	Mathematics		Science	
Years Since Treatment	**Impact on Boys' Test Scores**	**Impact on Girls' Test Scores**	**Impact on Boys' Test Scores**	**Impact on Girls' Test Scores**
1	0.22 (0.03) [0.000]	0.15 (0.03) [0.000]	0.18 (0.04) [0.000]	0.10 (0.05) [0.000]
2	0.35 (0.05) [0.000]	0.23 (0.05) [0.000]	0.26 (0.06) [0.000]	0.13 (0.06) [0.040]
3	0.42 (0.06) [0.000]	0.28 (0.06) [0.000]	0.29 (0.07) [0.000]	0.15 (0.07) [0.040]
4	0.47 (0.06) [0.000]	0.30 (0.07) [0.000]	0.30 (0.07) [0.000]	0.15 (0.07) [0.040]

Note: Each column shows estimates from the effects of one additional year of inquiry- and problem-based pedagogy after the number of years listed in the left column. Standard errors listed in parentheses and *p*-values listed in brackets. Standard errors and *p*-values are clustered by school.

Appendix: Supplemental Information

Trail Registry Information

(i) Paraguay 2011 and 2013 IPA IRB Protocol Number 241.11May-007 and AEA RCT Number [AEARCTR-0002947]

(ii) Peru 2012 Mathematics IPA IRB Protocol Number 12February-003 and 2014 IPA IRB Protocol Number:212.10April-002 for 2012, both with AEA RCT Number [AEARCTR-0000365]

(iii) Peru Science 2014 Mathematics IPA IRB Protocol Number 12February-003 and AEA RCT Number [AEARCTR-0000379]

(iv) Peru Science 2012 IPA IRB Protocol Number 215.10April-002 and AEA RCT Number [AEARCTR-0002960]

(iv) Belize [ISCR/H/2/71] and AEA RCT Number [AEARCTR-0002959].

The data for Argentina were provided to the authors by the government of Argentina, which executed implementation. Thus, we do not have registry or Institutional Review Board information.

Supplementary Table S1: Characteristics of IPP Interventions

Country/Year	Target Population	Grade	Didactic Materials	Teacher Training	Teacher Support	Source
Mathematics Interventions						
Argentina 2009	Public schools in Tafí Viejo, Yerba Buena, and Cruz Alta in Tucumán, and in southern Buenos Aires	4th grade	Workbook, calculator, rules, tables, games and figures	42 hours	Mentoring and training every other week	IDB (2018a)
Paraguay 2011	Preschools in Cordillera	Preschool	Workbook and audio lessons	35 hours	Mentoring and training once a month	IDB (2018c, 2018d)
Paraguay 2013	Preschools in Cordillera	Preschool	Workbook and audio lessons	35 hours	Mentoring and training once a month	IDB (2018e)
Peru 2012	Preschools in Huancavelica, Angaraes, and Ayacucho	Preschool	Mathematics tools (e.g., shapes, pictures, blocks, mirror, plastic tiles, and dice)	40 hours	Mentor visits once a month	IDB (2018f)
Belize 2015	Primary schools in Belize District	4th grade	Mathematics tools such as tin frames, geometric solids, rods, etc.	29 hours	Mentor visits once a month	IDB (2018b)
Science Interventions						
Argentina 2009	Public schools in socioeconomically disadvantaged communities in Tafí Viejo, Yerba Buena, and Cruz Alta in Tucumán, and in southern Buenos Aires	4th grade	Workbook and didactic materials	50 hours	Pedagogical and technical assistance	IDB (2018a)
Peru 2010	Public primary schools in Lima	3rd grade	LEGO kits	42 hours	Technical assistance and tutoring	IDB (2018g)
Peru 2012	Public primary schools in Lima	3rd grade	LEGO kits	73 hours	Technical assistance and tutoring	IDB (2018g)
Peru 2014	Students who perform in the bottom 50 percent on science scores in public primary schools in Lima	3rd grade	Flipcharts	20 hours	None	IDB (2018g)
Belize 2015	Primary schools in Belize District	4th grade	Mathematics tools such as tin frames, geometric solids, rods, etc.	29 hours	Mentor visits once a month	IDB (2018b)

Supplementary Table S2: Experimental Design Characteristics

Country/ Year	School Sample Frame	Number of Schools Sampled	Schools Allocated Treatment	Stratifications for Random Assignment	Baseline Collection Dates	Follow-up Collection Dates	Number of Students/ Baseline	Number of Students/ Follow-up
Argentina 2009	323	28	14	None	March 2009	November 2009	1,283	1,126
Paraguay 2011	265	265	131	Urban/rural, high/low school resources, and high/low school size	March 2011	November, December 2011	2,907	2,805
Paraguay 2013	265	262	129	Urban/rural, high/low school resources, high/low school size, and half sessions per day	March, April 2013	November 2013	3,195	2,888
Peru 2012	104	104	54	Urban/rural, and geographic department	March, April 2012	November 2012	2,926	2,400
Argentina 2009	323	42	28	None	March 2009	November 2009	2,271	1,927
Peru 2010	1203	106	53	Urban/rural/metro, complete/multigrade, and school size (small, medium, or large).	April 2010	December 2010	2,790	2,392
Peru 2012	1203	104	52	Urban/rural/metro, complete/multigrade, and school size (small, medium, or large)	March, April 2012	November, December 2012	2,705	2,401
Peru 2014	1217	48	Not applicable	School and gender	May 2014	November 2014	1,217	1,127
Belize 2015	258	252	25	Urban/rural and funding (government or government aided)	October, November 2014	May 2016	4,713	4,457

Supplementary Table S3: Definition of Variables Used in the Analysis

Variable	Definition
Panel A. Individual Characteristics	
Mathematics and science test scores (standard deviations)	Designed to measure the ability of students to understand and apply key mathematical and scientific concepts adapted for each grade level and national curriculum. Standardized to mean zero and standard deviation of 1 of the distribution of the control group.
Student's age	Age of student in years.
Male	Equals 1 if student is male and 0 otherwise.
Bilingual	Equals 1 if the child speaks Spanish and another language at home reported by parent and 0 otherwise.
Asset index (standard deviations)	Asset index created using principal component analysis to summarize information from the following variables: income per capita, number of people in the house, housing floor, ceiling, and wall materials. Standardized to mean zero and standard deviation of 1.
Panel B. School and Class Characteristics	
Average class size	Cohort size divided by number classrooms.
Teacher is male	Equals 1 if the sex of the teacher is male and 0 otherwise.
Teacher's age in years	Age of the teacher in years.

Supplementary Table S4: Baseline Descriptive Statistics and Tests of Balance between Treatment and Control Groups

	Mathematics					Science				
	Argentina 2009	**Belize 2015**	**Paraguay 2011**	**Paraguay 2013**	**Peru 2012**	**Argentina 2009**	**Belize 2015**	**Peru 2010**	**Peru 2012**	**Peru 2014**
Test scores	-0.02	0	0	-0.04	0.04	-0.02	0	-0.03	-0.03	-0.03
	(-0.07)	(-0.14)	(-0.02)	(-0.06)	(-0.11)	(0.01)	(0.01)	(0.09)	(0.03)	(-0.02)
	[0.267]	[0.202]	[0.719]	[0.416]	[0.284]	[0.739]	[0.959]	[0.434]	[0.746]	[0.737]
Age	9.35	8.26	5	4.9	5	9.36	8.26	n.a.	8.02	8.19
	(-0.05)	(0.04)	(0.01)	(0.00)	(0.00)	(0)	(0.04)		(-0.05)	(0.02)
	[0.17]	[0.63]	[0.249]	[0.971]	[0.545]	[0.95]	[0.63]		[0.34]	[0.699]
Male	0.52	**0.5**	0.5	0.53	**0.57**	0.52	**0.5**	0.52	0.52	0.55
	(-0.02)	**(0.07)**	(-0.02)	(0.00)	**(-0.08)**	(-0.04)	**(0.07)**	(0)	(-0.02)	(0.00)
	[0.497]	**[0.001]**	[0.165]	[0.896]	**[0.003]**	[0.284]	**[0.001]**	[0.943]	[0.347]	
Bilingual	n.a.	n.a.	0.43	n.a.	0.12	0.14	n.a.	n.a.	0.06	0.89
			(-0.01)		(0.00)	(0.01)			(-0.01)	(-0.01)
			[0.678]		[0.984]	[0.827]			[0.56]	[0.462]
Asset index	-0.04	-0.11	n.a.	n.a.	n.a.	-0.05	-0.11	n.a.	-0.01	n.a.
	(0.10)	(0.21)				(0.04)	(0.21)		(-0.04)	
	[0.074]	[0.1]				[0.76]	[0.1]		[0.573]	
Class size	**15.25**	23.42	15.36	17.13	21.45	**17.3**	23.42	23.36	22.48	23.81
	(-2.24)	(4.21)	(0.34)	(-0.13)	(2.02)	**(-2.05)**	(4.21)	(-1.52)	(-1.48)	(-0.06)
	[0.000]	[0.098]	[0.658]	[0.935]	[0.17]	**[0.083]**	[0.098]	[0.326]	[0.365]	[0.826]
Male teacher	n.a.	0.34	0.05	0.06	††	n.a.	0.34	n.a.	0.21	0.14
		(-0.05)	(0.04)	(0.02)			(-0.05)		(0.01)	(0)
		[0.712]	[0.205]	[0.586]			[0.712]		[0.901]	[0.748]
Teacher age	n.a.	35	35.25	37.14	n.a.	n.a.	35	n.a.	47.73	50.5
		(-2.59)	(0.75)	(0.04)			(-2.59)		(-0.16)	(-0.13)
		[0.278]	[0.314]	[0.953]			[0.278]		[0.927]	[0.741]

Note: The table shows the control group mean. The difference between the treatment and the control group means is shown in parentheses. The *p*-values for a test that the differences in means equals zero are shown in brackets. Errors are clustered at the school level, expect for those of Argentina, which are cluster bootstrapped. All estimates are based on baseline (pre-intervention) data. †† All teachers were female. n.a. denotes data not available. Bold font indicates statistically significant differences at the 0.10 significance level.

Supplementary Table S5: Attrition Rates between Baseline and Endline

Mathematics				Science				All
Argentina 2009	**Paraguay 2011**	**Paraguay 2013**	**Peru 2012**	**Argentina 2009**	**Peru 2010**	**Peru 2012**	**Peru 2014**	
0.13	0.03	0.08	0.21	0.17	0.16	0.11	0.08	0.08
(-0.01)	(0)	(0.02)	(-0.05)	**(-0.04)*** **	(-0.03)	(0.01)	(-0.01)	(-0.01)
[0.487]	[0.660]	[0.100]	[0.348]	[0.003]	[0.107]	[0.419]	[0.609]	[0.275]

Note: This table reports the attrition rate in the control group. Numbers in parentheses show the difference in attrition rates between the treatment and the control groups. The numbers in brackets show the corresponding *p*-values for the test that the difference in attrition rates equals zero with errors clustered at the school level. The standard errors for Argentina are cluster bootstrapped. *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.01 level.

Supplementary Table S6. Estimates of Equation (1) by Country, Year and Subject

	Mathematics					Science				
	Argentina 2009	**Belize 2015***	**Paraguay 2011**	**Paraguay 2013**	**Peru 2012**	**Argentina 2009**	**Belize 2015***	**Peru 2010**	**Peru 2012**	**Peru 2014**
β (Treatment)	0.13	0.16	0.20	0.17	0.19	0.08	0.29	0.17	0.14	0.12
	(0.06)	(0.09)	(0.05)	(0.04)	(0.06)	(0.04)	(0.09)	(0.08)	(0.08)	(0.05)
	[0.034]	[0.071]	[0.000]	[0.000]	[0.003]	[0.054]	[0.002]	[0.032]	[0.064]	[0.024]
γ (Lagged test score)	0.30		0.64	0.63	0.55	0.22		0.54	0.40	0.39
	(0.04)		(0.02)	(0.02)	(0.02)	(0.02)		(0.03)	(0.03)	(0.04)
	[0.000]		[0.000]	[0.000]	[0.000]	[0.000]		[0.000]	[0.000]	[0.000]
Sample Size										
Number of students	1,126	4,457	2,805	2,888	2,400	1,927	4,457	2,392	2,401	1,127
Number of schools	28	252	265	262	104	42	252	106	104	48

Note: Each column reports the estimates of the model in equation (1) for a different sample, including the estimated coefficients, standard errors in parentheses, and p-value for the hypothesis that the coefficient equals zero in brackets. The standard errors and p-values are clustered by school.
*Since we only have a cross-section for Belize 2015, we exclude the lagged test scores for those models.

Supplementary Table S7: Estimated Impacts on Test Scores and Cost-Effectiveness by Country and Subject

	Instantaneous Impact: One Year After Treatment		Long Run Impact: Four Years After Treatment	
	Impact on Test Scores	**U.S. Dollars per Student for a 0.10 Standard Deviation Increase in Test Scores**	**Impact on Test Scores**	**U.S. Dollars per Student for a 0.10 Standard Deviation Increase in Test Scores**
Mathematics				
Argentina 2009	0.13	US$5.84	0.19	US$3.99
Paraguay 2011	0.20	US$17.58	0.47	US$7.48
Paraguay 2013	0.17	US$22.48	0.39	US$9.22
Peru 2012	0.19	US$19.68	0.38	US$9.84
Science				
Argentina 2009	0.08	US$9.61	0.11	US$8.73
Peru 2010	017	US$17.52	0.33	US$9.56
Peru 2012	0.14	US$17.20	0.23	US$13.46
Peru 2014	0.12	US$49.96	0.20	US$29.97

Note: The cost to increase test scores by 0.10 standard deviations per student for Belize is 2015 was US$11.75.