IDB WORKING PAPER SERIES Nº IDB-WP-821

Do Tests Applied to Teachers Predict their Effectiveness?

Yyannú Cruz-Aguayo
Pablo Ibarrarán
Norbert Schady

Inter-American Development Bank
Social Sector - SCL/SCL

May, 2017

Do Tests Applied to Teachers Predict their Effectiveness?

Yyannú Cruz-Aguayo
Pablo Ibarrarán
Norbert Schady

May, 2017

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Cruz Aguayo, Yyannú.
Do tests applied to teachers predict their effectiveness? / Yyannú Cruz-Aguayo, Pablo Ibarraran, Norbert Schady.
p. cm. — (IDB Working Paper Series ; 821)
Includes bibliographic references.
1. Teacher effectiveness-Ecuador-Evaluation. 2. Teachers-Rating of-Ecuador. 3. Teaching-Ecuador-Evaluation. I. Ibarraran, Pablo. II. Schady, Norbert Rüdiger, 1967- III. Inter-American Development Bank. Education Division. IV. Title. V. Series.
IDB-WP-821

http://www.iadb.org

All three authors are with the Inter-American Development Bank.

Their e-mails are yyannuc@iadb.org, pibarraran@iadb.org, and norberts@iadb.org.

Do Tests Applied to Teachers Predict their Effectiveness?

Yyannú Cruz-Aguayo
Pablo Ibarrarán
Norbert Schady

May 8, 2017[1]

Abstract

Teachers vary considerably in their effectiveness, but identifying teacher characteristics that predict their impact on learning outcomes has been elusive. We analyze a teacher evaluation that is used to make teacher tenure decisions in Ecuador. The evaluation includes a written test, a demonstration class, and a points system that gives higher scores to teachers with more experience, degrees, and in-service training. We find no evidence that children taught by teachers with higher scores on the evaluation learn more. Our estimates are very precise: We can rule out that teachers with one-standard deviation higher evaluation scores raise child test scores in math by 0.03 standard deviations or more, and language scores by 0.02 standard deviations or more. We conclude that the effort that is being placed by policy-makers in Latin America to design and "improve" teacher tests is unlikely to result in large improvements in child learning.

JEL Codes: I25, I28

Keywords: Teachers, evaluation, test scores.

[1] All three authors are with the Inter-American Development Bank. Their e-mails are yyannuc@iadb.org, pibarraran@iadb.org, and norberts@iadb.org. We thank officials in the Ministry of Education in Ecuador for making data available to us, Fiorella Benedetti and Jorge Luis Castañeda for outstanding research assistance, and Pedro Carneiro for his comments.

1. Introduction

It is generally believed that teachers are the most important input into the production of learning within schools. Having a better teacher has long-term consequences for a variety of outcomes in adulthood, including college attendance, savings, and wages (Chetty et al. 2011; Chetty et al. 2014a). Research from the United States has credibly established that teachers—even teachers in the same school, teaching observationally equivalent students—vary a great deal in their effectiveness (Chetty et al. 2014b; Hanushek and Rivkin 2012). However, knowing that some teachers produce more learning than others provides no guidance on the attributes of effective teachers.

In this note, we analyze whether an evaluation that is used to determine which teachers receive tenure predicts student learning in Ecuador, a middle-income country in South America. In earlier work (Araujo et al. 2016), we showed that kindergarten teachers varied in the impact they have on learning: A one standard deviation increase in classroom quality increased test scores in math and language by 0.11 standard deviations.

In Ecuador, about one in three teachers in the public sector work on a contract basis. Tenured teachers are paid substantially more than contract teachers, and have more generous health insurance and pension benefits. Being converted from a contract to a tenured teacher therefore has important benefits for the individual, and implies increases in costs for the Ministry of Education.

Every year the Ministry carries out a planning exercise that takes account of the current distribution of tenured teachers, student population and the projected growth in demand, and any projected school openings, closings or merges. Based on this exercise, additional tenured slots are assigned to some schools. These slots are school-specific, but any individual who has a teaching degree from an accredited institution can apply. Among eligible applicants to a new slot, the applicant with the highest total score on an evaluation known as the *Concurso de Méritos y Oposiciones* (henceforth, *Concurso*) is awarded the tenured position.

The *Concurso* has three components, each of which receives equal weight in an aggregate score. The first component is given by a teacher's score on a test. This test has three parts: logical-verbal reasoning, pedagogical knowledge, and subject-specific knowledge. The second component of the aggregate score is a teacher's score on a demonstration class. The third component of the aggregate *Concurso* score is given by a point system (*Méritos*) which gives higher scores to

teachers with more experience, degrees (above and beyond the basic teaching certificate), and in-service training.

We use data on a sample of children in 2nd, 3rd, and 4th grades to analyze whether children taught by teachers with higher scores on the *Concurso* have higher achievement in language and math. Our earlier work (Araujo et al. 2016) was based on a sample of kindergarten children randomly assigned to teachers within schools. The data used for the analysis in this paper do not come from an experiment. We therefore have to make stronger identifying assumptions. In practice, we use a school fixed effects strategy that compares the test scores of observationally equivalent children taught by teachers with higher or lower *Concurso* scores in the same school.

2. Data and identification

Our sample consists of 4,479 children taught by 480 teachers in 240 schools. Schools were selected so each would have at least two contract teachers in 2nd through 4th grades who had taken the *Concurso* evaluation. Within each classroom, 9-10 children were randomly selected for testing. Twenty-eight percent of the children in the sample are in 2nd grade, 38 percent in 3rd grade, and 35 percent in 4th grade. To ease with interpretation, we standardize teacher scores on the *Concurso* and its components to have mean zero and unit standard deviation.

Children were tested twice, midway through the school year, and shortly before the end. We used adapted versions of the Early Grade Reading Assessment (EGRA; RTI International 2009; Gove and Cvelich 2011), and Early Grade Math Assessment (EGMA; Reubens 2009). Both tests have been applied in more than 20 developing countries. Application takes about 30 minutes per child. We standardize the four tests (EGRA and EGMA, mid- and end-year) so they have mean zero and unit standard deviation in each grade.

For 3,346 children in the sample (74.7 percent), we also have data from a household survey. This survey includes information on the years of completed schooling of mothers and fathers; whether a child attended preschool; and information on a large number of household characteristics (such as the material of the floors, wall, and ceilings in the house; whether the house is connected to the electricity, water, and sewerage system), and ownership of assets (questions for 23 assets). We aggregate the information on household characteristics and assets into a wealth composite (following, among many others, Araujo et al. 2016, and Paxson and Schady 2007, 2010, who also work with data from Ecuador).

The regressions we estimate take the following form:

$$(1)\ Y_{itgs} = \theta_g + \alpha_s + \beta_1 T_{gs} + \beta_2 X_{itgs} + \varepsilon_{itgs}, \qquad Z=1, 2, \ldots N$$

where the subscripts i, t, g, and s refer to individuals, teachers, grades, and schools, respectively; θ_g is a set of grade fixed effects; α_s is a set of school fixed effects; T_{gs} is the score of a teacher on the *Concurso* or its components; X_{itgs} always includes controls for child age in months, gender, and classroom size and, in some specifications, the education of mothers and fathers, household wealth, and an indicator variable for whether a child attended preschool; ε_{itgs} is the regression error term; and the regressions are run for z different tests (the mid- and end-year scores on EGRA or EGMA, and the change in scores between mid- and end-year). The coefficient of interest is β_1, an estimate of the extent to which children taught by teachers with higher scores on the *Concurso* have better scores in math or language. Standard errors are corrected for clustering at the school level.

The main estimation concern is purposeful sorting of children to teachers. As a check on our identification strategy, we run regressions of the variables in T_{gs} on the variables in X_{itgs}. Since the variables in X_{itgs} are predetermined, any indication that teachers with higher (or lower) test scores on the *Concurso* were assigned to children with different characteristics would raise questions about our identification strategy. Table 1 finds no evidence that this is the case, regardless whether the variables in X_{itgs} are entered one at a time or jointly. We conclude that our identification strategy is reasonable.

3. Results

To put things in context, we first regress the change in math or language test scores between mid- and end-year on the full set of child and household controls. We take the residuals from these regressions, calculate classroom means, and the difference in means between classrooms in the same school. Within-school differences are large: The median difference is .18 standard deviations for language, and .22 standard deviations for math. Although some of these within-school, cross-classroom differences are probably driven by sampling error or idiosyncratic classroom shocks, it seems likely that a substantial proportion reflect true differences in the quality of teachers.[2]

[2] In Araujo et al. (2016) and Carneiro et al. (2017) we use an Empirical Bayes estimator to shrink the estimated classroom effects. On average, the corrected estimates are about one-third smaller than those that do not correct for sampling error. We do not do this in the current paper because we are not calculating classroom (or teacher) effects.

Next, we show that the characteristics in X_{itgs} tend to be strongly correlated with test scores, as expected. Table 2 shows that girls score about 0.14 standard deviations higher on language, and 0.23 standard deviations lower in math.[3] Children from households of higher socioeconomic status have higher test scores in both reading and math.

Our main findings are in Table 3. We report the results from specifications with the additional controls (even-numbered columns) and without (odd-numbered columns), and six different dependent variables: the mid- and end-year test scores in language and math, and the change in test scores between mid- and end-year.[4]

Table 3 shows no evidence that the test score or any of its components, the score on the demonstration class, points on the *Méritos* scale, or the aggregate score on the *Concurso* predict child achievement in language or math. There are 84 coefficients in the table, and about one-half (40) are positive and one-half (44) are negative. No coefficient is significant at conventional levels. The estimates in the table are quite precise. For example, in the specifications with the additional controls, we can rule out positive associations between the total *Concurso* score and child test scores larger than 0.02 standard deviations for language, and 0.03 standard deviations for math. Consistent with the results in Table 1, the coefficients with and without the additional controls are very similar.

In sum, although there are large differences across classrooms in the same school in how much children learn, there is no evidence that children taught by teachers with better *Concurso* scores are more effective. If teachers with better scores were systematically assigned students with lower learning potential, our results could be biased towards zero. We find no evidence that this is the case. We conclude that the evaluation that was used to make tenure decisions in Ecuador did not predict how effective teachers were at increasing the language and math test scores of children in elementary school.

[3] Carneiro et al. (2017) carefully analyze the gender gap in math achievement in kindergarten through 2nd grade in Ecuador. Consistent with what we report here, they find a gender gap of 0.17 standard deviations, favoring boys, by 2nd grade.

[4] Results are very similar if instead we use the largest possible sample of children and do not include the additional controls. These results are available from the authors upon request.

4. **Discussion and conclusion**

Improving the quality of education is an important priority for Latin American countries. Teachers vary considerably in their effectiveness and many Latin American children are taught by low-quality teachers (Bruns and Luque 2015). The region has low productivity, much of which can be traced to the low skills of the workforce (Pages 2010). Some analysts have argued that the low learning outcomes of Latin American students (and eventually adults) can account for much of the difference in growth rates between Latin America and East Asia (Hanushek and Woessman 2012).

Research from the United States has generally been pessimistic about the ability of various tests to identify better teachers (Hanushek and Rivkin 2006; Staiger and Rockoff 2010). We carefully analyze a teacher evaluation that included a written test, a demonstration class, and points for experience, degrees, and in-service training. Earlier work (Araujo et al. 2016) shows that teachers in Ecuador vary considerably in the impact they have on learning outcomes. In this note we show, however, that the instrument that is used to decide which teachers get tenure in Ecuador—an instrument similar to that used in many other Latin American countries—does not predict how effective a teacher is at raising math and language achievement.[5] Our estimates are quite precise and we can rule out even very modest effects on test scores. The results in our paper, and others in the literature, lead us to be cautious about what one can, and cannot, expect from a teacher testing regime like those that have become popular in many developing countries.[6]

[5] Many countries in Latin America, including Chile, Colombia, El Salvador, Mexico, Peru, and some municipalities and states in Brazil, test teachers. The tests generally include a written component and a demonstration class. Performance on this teacher evaluation is an input into decisions about hiring, tenure, in-service training, pay and, in some countries (like Chile), which teachers can get fired.

[6] An influential report on teachers in Latin America concludes that "mandatory certification exams are the most powerful instrument for raising teacher standards" (Bruns and Luque 2015, p. 31). Our results are not consistent with this assertion.

References

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-53.

Bruns, Barbara, and Javier Luque. 2015. *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean*. Washington, D.C.: The World Bank.

Carneiro, Pedro, Yyannu Cruz-Aguayo, and Norbert Schady. 2017. "Where the Girls Are Not: Households, Teachers, and the Gender Gap in Early Math Achievement." Unpublished manuscript, Inter-American Development Bank.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. ''How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR.'' *Quarterly Journal of Economics* 126(4): 1593–1660.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014a ''Measuring the Impact of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593–632.

———. 2014b. ''Measuring the Impact of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.'' *American Economic Review* 104(9): 2633–679.

Gove, Amber, and Peter Cvelich. 2011. *Early Reading: Igniting Education for All. A Report by the Early Grade Learning Community of Practice* (Revised Edition). Research Triangle Park, N.C.: Research Triangle Institute.

Hanushek, Eric, and Steven Rivkin. 2006. "Teacher Quality." In Eric Hanushek and Finis Welch, Eds., *Handbook of the Economics of Education*, Vol. 2. Elsevier.

----------. 2012. "The Distribution of Teacher Quality and Implications for Policy." *Annual Review of Economics* 4: 131–57.

Hanushek, Eric A., and Ludger Woessmann. 2012."Schooling, Educational Achievement, and the Latin American Growth Puzzle." *Journal of Development Economics* 99(2): 497–512.

Pages, Carmen. 2010. *The Age of Productivity*. New York: Palgrave Macmillan.

Paxson, Christina, and Norbert Schady. 2007. "Cognitive Development among Young Children in Ecuador: The Roles of Wealth, Health, and Parenting." *Journal of Human Resources* 42(1): 49-84.

----------. 2010. "Does Money Matter? The Effects of Cash Transfers on Child Health and Development in Rural Ecuador." *Economic Development and Cultural Change* 59(1): 187-229.

Reubens, Andrea. 2009. *Early Grade Mathematics Assessment (EGMA): A Conceptual Framework Based on Mathematics Skills Development in Children*. Research Triangle

Park, NC: Research Triangle Institute.

Staiger, Douglas O., and Jonah E. Rockoff. 2010. “Searching for Effective Teachers with Imperfect Information.” *Journal of Economic Perspectives* 24(3): 97-118.

Table 1: Associations between teacher scores on the *Concurso* and predetermined child and household characteristics

	CHILD OR HOUSEHOLD CHARACTERISTIC							
TEACHER SCORE	Age (months)	Female	Class Size	Mother Schooling	Father Schooling	Wealth	Attended Preschool	F-test Prob > F
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test score: aggregate score	.002	-.014	-.008	.001	.003	-.026	-.027	0.31
	(.004)	(.018)	(.015)	(.003)	(.003)	(.024)	(.040)	
Test score: verbal-logical reasoning	-.004	.008	-.008	.004	.005	-.010	.025	0.50
	(.004)	(.021)	(.016)	(.003)	(.003)	(.019)	(.036)	
Test-score: pedagogical practices	-.004	-.009	-.006	.002	-.001	-.020	.022	0.93
	(.005)	(.016)	(.013)	(.003)	(.003)	(.032)	(.044)	
Test score: content	.003	-.027	-.005	-.003	.003	-.027	.103**	0.07
	(.004)	(.023)	(.018)	(.003)	(.004)	(.017)	(.041)	
Demonstration class	-.000	.019	-.004	.009**	.006	.033	.023	0.13
	(.004)	(.031)	(.013)	(.004)	(.004)	(.024)	(.043)	
Points on *Méritos* scale	.002	-.023	-.008	.003	-.002	.014	-.011	0.73
	(.003)	(.020)	(.014)	(.003)	(.003)	(.017)	(.039)	
Total *Concurso* score	.002	-.030	.007	.004	.002	.014	-.010	0.85
	(.004)	(.024)	(.014)	(.004)	(.003)	(.017)	(.033)	

Notes: Regressions of teacher scores on characteristics in columns (1) through (7), with each teacher score regressed on each child or household characteristic (49 separate regressions). F-test in column (8) refers to an F-test from a regression in which all the characteristics in columns (1) through (7) enter jointly. All regressions include school fixed effects. Teacher scores have been transformed to have mean zero and unit standard deviation. Sample size is 3,058 in every regression. Standard errors clustered at the school level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent level, respectively.

Table 2: Associations between child and household characteristics, and child test scores

	CHILD OR HOUSEHOLD CHARACTERISTIC						
CHILD TEST SCORE	Age (months)	Female	Class Size	Mother Schooling	Father Schooling	Wealth	Attended Preschool
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
EGRA: mid-year	.008**	.141***	-.007	.040***	.036***	.134***	.010
	(.003)	(.032)	(.006)	(.004)	(.005)	(.024)	(.045)
EGRA: end-year	.007**	.135***	-.003	.045***	.036***	.139***	.073*
	(.003)	(.032)	(.006)	(.004)	(.005)	(.024)	(.043)
EGMA: mid-year	-.004	-.242***	-.006	.039***	.032***	.112***	-.003**
	(.003)	(.034)	(.005)	(.005)	(.005)	(.027)	(.045)
EGMA: end-year	-.002	-.223***	-.002	.043***	.030***	.132***	.078*
	(.003)	(.033)	(.006)	(.005)	(.005)	(.024)	(.047)

Notes: Regressions of child test scores on characteristics in columns (1) through (7), with each child score regressed on each child or household characteristic (28 separate regressions). All regressions include school fixed effects. Child test scores have been transformed to have mean zero and unit standard deviation. Sample size is 3,322 in every regression. Standard errors clustered at the school level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent level, respectively.

Table 3: Main results

	TEACHER SCORE ON CONCURSO OR COMPONENTS													
	Total test score		Verbal-logical reasoning score		Pedagogical practices score		Content score		Demonstration class		Points on *Méritos* scale		Total *Concurso* score	
CHILD TEST SCORE	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
EGRA: mid-year	.020	.018	.040	.032	-.002	-.002	.014	.014	.002	-.007	-.021	-.021	.000	-.005
	(.037)	(.037)	(.037)	(.037)	(.029)	(.028)	(.042)	(.043)	(.023)	(.023)	(.034)	(.034)	(.030)	(.029)
EGRA: end-year	.021	.020	.030	.023	.000	.000	.020	.021	.013	.003	-.032	-.033	-.004	-.010
	(.035)	(.034)	(.037)	(.037)	(.028)	(.027)	(.041)	(.041)	(.023)	(.023)	(.033)	(.033)	(.029)	(.028)
Δ EGRA score	.001	.001	-.009	-.010	.003	.002	.006	.008	.011	.010	-.011	-.011	-.004	-.004
	(.018)	(.018)	(.016)	(.016)	(.016)	(.016)	(.015)	(.014)	(.009)	(.009)	(.015)	(.015)	(.012)	(.013)
EGMA: mid-year	-.046	-.048	-.026	-.033	-.042	-.042	-.026	-.025	.021	.012	.026	.025	-.003	-.008
	(.031)	(.031)	(.036)	(.036)	(.027)	(.027)	(.034)	(.034)	(.025)	(.024)	(.029)	(.029)	(.029)	(.027)
EGMA: end-year	-.046	-.046	-.056	-.062	-.022	-.022	-.027	-.022	.018	.008	.042	.040	-.012	-.018
	(.033)	(.033)	(.036)	(.037)	(.029)	(.030)	(.034)	(.034)	(.022)	(.021)	(.031)	(.030)	(.029)	(.027)
Δ EGMA score	-.000	.002	-.030	-.030	.021	.020	-.001	.003	-.002	-.003	.016	.016	-.009	-.009
	(.018)	(.018)	(.020)	(.020)	(.015)	(.015)	(.017)	(.017)	(.012)	(.012)	(.018)	(.018)	(.017)	(.017)

Notes: Regressions of child test scores (mid-year, end-year, or changes between mid-year and end-year) on teacher scores on the *Concurso* or its components (84 separate regressions). All regressions include school fixed effects, class size, child age in months, and child gender. Specifications in even-numbered columns also include education of mothers, fathers, household wealth, and whether the child attended preschool. Child test scores and teacher scores on the *Concurso* have been transformed to have mean zero and unit standard deviation. Sample size is 3,058 in every regression. Standard errors clustered at the school level. *, **, and *** indicate significance at the 10 percent, 5 percent, and 1 percent level, respectively.