IDB WORKING PAPER SERIES No. IDB-WP-529

The Effect of In-Service Teacher Training on Student Learning of English as a Second Language

Rosangela Bando
Xia Li

July 2014

Inter-American Development Bank
Office of Strategic Planning and Development Effectiveness

The Effect of In-Service Teacher Training on Student Learning of English as a Second Language

Rosangela Bando
Xia Li

Inter-American Development Bank
2014

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Bando, Rosangela.
The effect of in-service teacher training on student learning of English as a second language / Rosangela Bando, Xia Li.
p. cm. — (IDB Working Paper Series ; 529)
Includes bibliographic references.
1. Teachers—Training of—Mexico. 2. Teachers—In-service training —Mexico. 3. English language—Study and teaching—Foreign speakers. I. Li, Xia. II. Inter-American Development Bank. Office of Strategic Planning and Development Effectiveness. III. Title. IV. Series.
IDB-WP-529

http://www.iadb.org

1300 New York Avenue, NW, Washington, DC 20577

Rosangela Bando, rosangelab@iadb.org

The Effect of In-Service Teacher Training on Student Learning of English as a Second Language[1]

Rosangela Bando and Xia Li[2]

Inter-American Development Bank

Draft version: July 24, 2014

Abstract

In-service teacher training aims to improve the supply of public education. A randomized experiment was conducted in Mexico to test whether teacher training could increase teacher efficiency in public secondary schools. After seven and a half months of exposure to a trained teacher, students improved their English. This paper explores two mechanisms through which training can affect student learning. First, trained teachers improved their English by 0.35 standard deviations in the short run. Teachers in the control group caught up with treatment teachers by the end of the school year in part because teachers in the treatment group reduced out-of-pocket expenditures to learn English in 53 percent. Second, teachers changed classroom practices by providing more opportunities for students to actively engage in learning. This evidence suggests that teacher training may be effective at improving student learning and that teacher incentives may play a role in mediating its effects.

JEL classification: I21, O15, M53

Keywords: Human capital formation, analysis of education, teacher training, English as a second language, secondary education.

[1]We are indebted to the Worldfund team and the Ministry of Education in Puebla and Tlaxcala, who provided logistical support and provided the necessary information to make the evaluation possible: Mary Bourque, Jim Citron, Juan Odín Cano, Helene Rassias, Isabel Reyes, and Patricia Vazquez. Sebastián Galiani provided useful comments. Carola Alvarez and Marcelo Cabrol provided insights and full institutional support for the evaluation. Steve Marban, Raúl Abreu, and Armando Loera provided support to make data collection possible. Aniel Altamirano provided excellent research assistance. 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, or the countries they represent.

[2]Corresponding author: Rosangela Bando. 1300 New York Avenue, NW, Washington, DC 20577. Phone: (202) 623-2126. Email: rosangelab@iadb.org

1 Introduction

This study explores the effects of an in-service teacher training program on student performance in Mexico. Random allocation of 77 teachers to a treatment group and 67 to a control group allow for identification of causal effects on student learning as measured by standardized test scores. The study was carried out on a sample of teachers teaching English as a second language (ESL) in public secondary schools in the states of Puebla and Tlaxcala, Mexico. Training provided to treatment teachers aims to enhance ESL acquisition in students by changing classroom practices and increasing English teachers' subject matter knowledge. A comparison between students of teachers in the treatment group and students of teachers in the control group shows that teacher training improves student performance. The students of trained teachers improved by around 0.16 standard deviations when compared to students of non-trained teachers in an average of seven and a half months of exposure.

This paper explores two mechanisms through which training can affect student learning: changes in teacher subject matter knowledge and changes in classroom practices. The main contribution of this paper is that it provides quantitative evidence of whether or not in-service teacher training alone can change teacher and student behavior to improve student learning through a randomized controlled trial and its underlying mechanisms.

Providing such evidence is relevant for three reasons. First, considerable resources are increasingly being devoted to teacher training. For example, the federal budget for the Office of Professional Development of Teachers in Mexico increased 7.7 times from 2005 to 2010, with over US$370 million allocated in 2010 (INEGI, 2012). However, evidence of the effects of in-service teacher training on student learning is limited. The Ministry of Education in Mexico was concern about the

lack of evidence and reported in 2003 that investment in in-service teacher training, or professional development, has likely not had much of an effect on student performance (Presidencia de la República, 2013).

Investment in in-service teacher training is motivated by qualitative evidence and international comparisons across educational systems which support the idea that teacher training should lead to student learning (Wayne et al., 2008; Vegas and Petrow, 2008). One limitation of qualitative studies and international comparisons is that it is difficult to disentangle the effects of teacher training from differences between specific units in the study, other inputs, and the context.

International quantitative evidence in the past has focused on disentangling the effects of in-service teacher training and school and individual teacher characteristics. Teachers that are better trained and more experienced tend to seek work in schools with students with more ability (Harris and Sass, 2011). There are usually strong self-selection effects, and this tends to lead non-experimental designs to overestimate effects (Mano et al., 2013; Hamalainen, Uusitalo, and Vuori, 2008; Hotz, Imbens, and Mortimer, 2005). Harris and Sass (2011) review quantitative studies aiming to assess the effects of teacher training on student learning. The authors find only three randomized controlled trials on the effect of in-service teacher training on student learning, all carried out in the United States. This study is one of the few that addresses the self-selection issue and provides a causal estimate of the effect of in-service teacher training on student learning.

Second, a better understanding of how in-service teacher training affects teacher behavior is needed in order to improve future program implementation. Teacher training is a priority in Mexico and in many other countries. The current development plan in Mexico has the explicit goal of boosting in-service teacher training

to enhance teachers' understanding of the educational system and improve pedagogical practices and management of information technologies (Presidencia de la República, 2013). However, international quantitative evidence on the effects of in-service teacher training on student performance is mixed. Several studies conclude that there are likely no effects on student learning (Metzler and Woessmann, 2010; Chingos and Peterson, 2011; Garet et al., 2011; Harris and Sass, 2011; Rockoff, 2004; Carnoy et al., 2008; Jacob and Lefgren, 2004). Other studies support the idea that training has the potential to increase productivity (Hsiao et al., 2008; Sunardi, Widyarini, and Tjakraatmadja, 2012; Angrist and Lavy, 2001; Mason, O'Leary, and Vecchi, 2012).

Most studies emphasize the role of context and complementarities in bringing about the observed results. Thus, it is important to better understand how teacher training affects classroom practices and teacher behavior. Qualitative evidence supports the idea that in order for in-service teacher training to be effective it has to have some specific characteristics, such as being connected to practice, intensive enough, linked to incentives, and continuous (Darling-Hammond et al., 2009). This study explores underlying mechanisms to better understand how in-service teacher training works.

This paper provides evidence of how teacher training changes teacher subject matter knowledge and classroom practices. It is closely related to Goldschmidt and Phelps (2010a), who examine the impact of teacher professional development on teacher knowledge. Their results indicate that teachers demonstrate significant knowledge growth between the pre- and post-assessments but that practical classroom experience hinders knowledge retention, measured six months later.

The results of this study are similar in that teachers improve their English subject

matter knowledge right after the training, but they tend to forget it in the long run. However, we show that it is the reduction of personal investment in learning English by the teachers in the treatment group that hinders knowledge retention in the long run. Instead, teachers improve their classroom practice by encouraging effective and active learning. This is very different from Goldschmidt and Phelps (2010a)'s finding. Also, this analysis looks at a developing country, which could be very different from what occurs in the United States, as in Goldschmidt and Phelps (2010a). This study is also closely related to that of Muralidharan and Sundararaman (2010), who find that teacher incentives mediate the effects of teacher training in India. This study is also one of the few that isolate the effects of teacher training from changes in other inputs or factors that usually accompany training, such as changes to curricula, provision of didactic materials, or technology. This study is one of the first to identify the causal effect of in-service teacher training alone on student learning in a developing country.

Third, studying English acquisition is important because the ability to communicate in English determines the extent in which an individual can participate in the global labor market. Estimates of returns to English acquisition in countries other than Mexico and in immigrants in the United States vary between 4 and 39 percent, with stronger effects in more educated individuals (Azam, Chin, and Prakash, 2013; Gonzalez, 2005; Bleakley and Chin, 2004; Munshi and Rosenzweig, 2006).

Current investment in English acquisition is large and increasing. The current curriculum devotes 11 percent of class time to English in high school, and the government plans to make it compulsory at the pre-primary and primary levels in the future. Focusing on public education is important. Eighty-seven percent of students in basic education are enrolled in public schools. Estimates show that 80 percent of

the population cannot afford private English instruction. He, Linden, and MacLeod (2008)conducted one of the few studies that specifically addresses English acquisition.

The remainder of the paper is organized as follows. Section 2 contains a description of the program and the background. Section 3 describes the sample, the data, and the evaluation design. Section 4 describes the estimation of program impacts on student learning. Section 5 explores the underlying mechanisms through which training leads to learning. Section 6 provides robustness checks to eliminate the possibility of potential threats to estimation. Section 7 provides a short summary of results and discusses policy implications.

2 Background and Program Description

This study evaluated the impact of a ten-day component of the Inter-American Partnership for Education (IAPE) program, a Clinton Global Initiative Commitment that aims to empower classroom teachers in Mexican public schools with the English language and teaching skills to effectively inspire students to master the core English competencies established by their state or national authorities.[3] The program is provided at a cost of MXN$22,500 per teacher.[4] The program targets teachers with an English level comparable to an approximate TOEFL between 320 and 499 points or levels A1+, A2 or B1 according to the Common European Framework of Reference[5]. The teacher selection process is a joint effort of Dartmouth College's Rassias Center, Worldfund, and Mexican state governments. In a first

[3]Source: IAPE webpage http://worldfund.org/index.php/our-programs/iape as retrieved on April 30, 2012.

[4]In 2013 Mexican pesos.

[5]Teachers with an English proficiency score of over 499 are advised to apply to another program called the IAPE Teachers' Collaborative program, which takes place at Dartmouth College.

stage, teachers are invited by the state to voluntarily participate and apply online. In a second stage, IAPE-trained instructors conduct an eight- to ten-minute phone interview to assess the teacher's English proficiency. The score must fall between 3 and 7 on a 0 to 10 scale, which is approximately equivalent to the levels targeted by the program.

The program provides 100 hours of intensive training, 80 of which are devoted to intensive English instruction and 20 hours to pedagogical training. The program requires teachers to speak only English during their course of study. Teachers are encouraged to use only English as the language of instruction in their own practices. The program encourages teachers to use dynamic activities sustained in repeated interactions under a rhythm. These activities aim to increase class participation. The program also encourages the use of humor and drama in class in order to keep students engaged and speaking. The program provides techniques to encourage teachers and students to remove inhibitions about speaking English. The program is designed to offer periodic follow-up by encouraging trained teachers with higher levels of English proficiency to create virtual communities where both trained and non-trained teachers receive support. The follow-up component of the program was not implemented in the evaluation sample because contamination into control teachers would have been unavoidable and there was no option to create "pure controls" out of the evaluation sample.

Currently, the program is not aligned with any teacher incentives. The role of teacher incentives in in-service teacher training is relevant because they affect participation and effectiveness of teacher training programs. Muralidharan and Sundararaman (2010) studied the effects of teacher incentives on student performance in India and found that teacher incentives were especially effective among teachers

with more training but that training alone was not a predictor of student improvements. In this study, the lack of teacher incentives resulted in only about 50 percent of eligible teachers applying for the program. The states covered all of the participants' expenses and provided all of the requisite permits to attend the training. The lack of teacher incentives also contributed to the substitution of private investment in English language learning.

3 Sample, Evaluation Design and Data

3.1 Sample

The evaluation sample consists of 144 teachers actively working at the secondary level (grades 7 to 12) in public schools in the states of Puebla and Tlaxcala. External validity for this experiment is relevant for teachers that qualify for the program and is limited within the states. Teachers in the sample were recruited in phases. Some teachers attritted from the sample.

The recruitment process began with an invitation sent to 439 teachers in the state of Puebla who had voluntarily participated in an English language training program at the local state university, the Benemérita Universidad del Estado de Puebla (BUAP), and were assessed to be within the English proficiency range to benefit from the program. In a first cohort, 78 teachers were identified as eligible to be trained May 14-24, 2012. In a second cohort, 63 teachers were identified as eligible to be trained September 20-30, 2012. For a third and final cohort, the invitation was extended to the neighboring state of Tlaxcala, which identified 25 qualified teachers in lower secondary schools. For the third cohort, 64 teachers were identified as eligible in the state of Puebla. Of the 64 teachers, 25 teach in lower secondary schools in Tlaxcala. Teachers identified as eligible in the third round were trained

November 7-18, 2012. In total, 205 teachers were recruited for the three cohorts, out of which 182 were confirmed to be English teachers actively teaching ESL in a secondary public school. Twenty-one teachers were lost due to attrition. A group of 17 teachers is excluded from the main analysis because attrition was correlated to treatment in this group. The evaluation sample consists of the 144 remaining teachers.

Teachers across cohorts may not be comparable. Teachers in the first cohort may be more motivated to receive training, may have better communication with the state, or may have greater English proficiency.

The selection process affects the external validity of the results. The degree of English proficiency of participating teachers is very likely higher than that of non-participating teachers. Forty-seven percent of the teachers who were assessed scored too low to participate, and only 4 percent scored too high. Participating teachers are located in large urban areas. One-third of the teachers were assessed in municipalities outside of the capital, and none of them scored higher than the lowest rate (A1) and therefore were not eligible for the program. Participating teachers are more likely to be female, teach in larger schools close to or in the capital of the state and in localities with lower poverty rates. Results are valid on teachers with similar English proficiency and in contexts similar to those of Puebla and Tlaxcala. This is likely to be the case in states such as Tamaulipas, Nuevo León, Aguascalientes, Morelos, Coahuila, Sinaloa, Sonora, Colima, Baja California Sur, and the Distrito Federal. These states have made progress in English training for teachers above the national average, according to the Ministry of Education.[6].

[6] Programa Nacional de Educación Básica, http://basica.sep.gob.mx/pnieb/start.php?act=mapaEstados. Retrieved on October 13, 2013.

3.2 Data

Data were collected in two rounds: baseline and follow-up. Baseline data were collected on the English proficiency and general characteristics of the teachers before training sessions for each cohort started with the Diagnostic Instrument for the Measurement of English (DIME) test and a closed question questionnaire. The DIME test was developed using best practices in test item development and reflects rigorous English language proficiency measurement protocols. DIME was originally developed to assess English teachers in the Mexican Ministry of Education and has been validated for the purposes of this study.[7] The DIME test presents 45 multiple-choice questions, five fill-in-the-blank questions, and requires a writing sample and a recorded speaking sample. The total DIME test score ranges from 0 to 120 points and the listening, writing, speaking, and reading sections have scores that range from 0 to 30 points. For treatment teachers, the DIME test is also applied right after training to assess effects in the short run.

Follow-up data were collected at the end of the school year, in May to June of 2013. For this round, the DIME and closed questionnaire for teachers applied at baseline are collected again. In addition, one class per teacher is observed and recorded on video. The class observation is made using the Stallings instrument (Stallings, 1980). The Stallings instrument records classroom activities, materials used, and class organization for each tenth of the class length. Stallings allows time allocation in the classroom to be estimated by analyzing activities in the snapshots. The recorded videos are coded according to the applicable dimensions in the TIMSS

[7]The research base used in the development of the test is that used in the development of the ELPA test developed by the University of Michigan, the TOEIC Bridge test developed by the Educational Testing Service, and the battery of tests developed by Cambridge English. The validation was against the Stanford battery of tests published by NCS Pearson.

1999 video study (National Center for Education Statistics, 2003). The follow-up round also collected information on students. Five randomly selected students per teacher were tested with a DIME English test specifically designed for students. Students were also asked to answer a closed question questionnaire to obtain information on general characteristics, investment, and attitudes toward learning English. The resulting dataset allows for baseline and follow-up information on the English proficiency of 144 teachers. Information for students and classroom dynamics is only available at follow-up.

3.3 Evaluation Design

One of the main contributions of this study is to establish the causal effect of teacher training on student learning. High-performing teachers tend to seek to teach in high-performing schools. As a result, a comparison of student learning in teachers that choose to receive English language training would likely incorporate differences in teacher characteristics, such as motivation, and student characteristics, such as initial performance. The 205 eligible teachers identified before school visits were randomly allocated to receive training or not. Within each of the three cohorts recruited, teachers were stratified by whether they reported teaching in lower or upper secondary schools and by state. Within each stratum, teachers were randomly allocated either to treatment or control groups. Random allocation creates two groups of teachers with students that are not different before program intervention. As a result, differences in student learning and teacher behavior observed after training can be attributed to teacher training.

One threat to identification is that the exclusion of the 23 school personnel miss reporting as English teachers and the 21 teachers lost to attrition may change the composition of either the treatment or the control group. We discuss the threat of

the 23 miss reporting first and the 21 teachers lost to attrition second.

First, it is unlikely that miss reporting could change the composition of the treatment or the control groups. Randomization took place after the teachers had declared a teaching status and proficiency level and therefore could not be correlated to treatment. We test this empirically and find that miss reporting is not correlated in all but the stratum corresponding to the second wave upper secondary level in Puebla. This cohort is composed of 23 individuals where 17 are confirmed as English teachers. This cohort is excluded from the study but results are robust to its inclusion.[8]. The final sample consists of 165 teachers in the six remaining strata. Table 1 shows the number of teachers allocated to treatment and control groups by wave and proficiency level.

Tests of differences between treatment and control groups are made to determine whether randomization was successful at creating two groups that are not different on average before program intervention. Table 2 shows that there are no differences in the teachers' English proficiency levels, individual characteristics, or school characteristics. Imbens and Wooldridge (2009) argue that when normalized differences exceed one quarter, then linear regression methods tend to be sensitive to the specification. The table reports standardized differences.

Second, we discuss attrition of 21 teachers. Attrition can be problematic if teachers with specific characteristics leave the sample in either group. For example, if better-performing teachers are less likely to leave the treatment group because of program participation, then ignoring attrition would lead to a comparison of two groups that differ both in program participation and teacher composition. Table

[8]For this cohort of 23 individuals, six are not English teachers, of which four are in the control group and two in the treatment group. Out of the 17 teachers in the cohort, two more in the control group and one in the treatment group are lost to attrition. Therefore this group is excluded, as potential differences in attrition patterns may threaten identification.

Table 1: Individuals by Stratum Allocation

	1st cohort		2nd cohort		3rd cohort		
School type	Treatment	Control	Treatment	Control	Treatment	Control	Total
Recruited individuals over which randomization was carried							
Lower secondary - Puebla	29	30	21	19	3	3	105
Upper secondary - Puebla	11	8	11*	12*	17	16	75
Lower secondary - Tlaxcala					13	12	25
Total	40	38	32	31	33	31	205
Recruited teachers							
Lower secondary	25	26	18	20	16	15	120
Upper secondary	10	8	9*	8*	15	12	45
Total	35	34	18	20	31	27	165
Teachers with follow up information							
Lower secondary	24	22	18	19	14	12	109
Upper secondary	7	6	8*	6*	14	8	35
Total	31	28	18	19	28	20	144

The first stratum is composed of the 59 teachers in the lower secondary level in Puebla for the first cohort, the second stratum is composed of the 19 teachers in the upper secondary level in Puebla for the first cohort, the third stratum is composed of the 40 teachers in lower secondary level in Puebla for the second cohort and so on. * This stratum is excluded from the analysis because attrition may bias results. Results are robust to its inclusion in the analysis.

3 shows a comparison of the teachers that remained in the sample and those that left. Attrition was higher among teachers in the control group. Other than program participation, there are no differences in other observable characteristics including English proficiency. The differences in attrition between treatment and control groups do not hold once teachers are controlled by their socioeconomic status.[9]

Table 4 shows conditional and unconditional balance on observable teacher characteristics on the sample of 144 teachers for which there is information to evaluate. In the evaluation sample, the treatment and control groups are equal on average be-

[9]The P-value for equality in means in attrition rates is 0.221 when controlled for the socioeconomic status. The SES of teachers is a dummy which equals one if the socioeconomic index is above the median and zero otherwise. The socioeconomic index is created using principal component analysis including household assets and dwelling characteristics. Dwelling characteristics include number of rooms, bathrooms, internet access and availability of a quiet place to work. Assets include DVD or VCR, cable, microwave oven, cell phone, television, computer, car, other vehicles, desk for work, encyclopedias or dictionaries, and English courses. Results are robust to the inclusion of a dummy or the index.

Table 2: Baseline Means over Randomization Sample

	Treatment	Control	Standardized differences in means*	P-value (value for test of equality)
English level				
DIME total	66.68	63.61	-0.16	0.162
	(1.66)	(1.41)		
Writing	20.62	20.03	-0.08	0.449
	(0.54)	(0.56)		
Speaking	18.39	18.25	-0.02	0.842
	(0.51)	(0.50)		
Listening	14.81	13.03	-0.28	0.013
	(0.47)	(0.52)		
Reading	12.86	12.30	-0.08	0.459
	(0.56)	(0.52)		
CEFR (1 if B1 and 0 if A1+ or A2)	0.38	0.46	0.11	0.326
	(0.05)	(0.06)		
Oral interview score (scale 0 to 10)	4.78	4.82	0.02	0.852
	(0.15)	(0.15)		
School characteristics				
Teaches in lower secondary school	0.70	0.75	0.08	0.468
(1 if yes 0 if no)	(0.05)	(0.05)		
School size (number of students)	322.38	336.67	0.04	0.711
	(26.49)	(27.96)		
Individual characteristics				
Gender (1 if male, 0 if female)	0.35	0.35	-0.00	0.982
	(0.05)	(0.05)		
Age (in years)	37.55	38.00	0.04	0.733
	(0.84)	(0.99)		
Education above teacher certification	0.80	0.73	-0.13	0.231
(1 if yes, 0 if no)	(0.04)	(0.05)		
Currently studying (=1 if yes, 0 if not)	0.37	0.44	0.10	0.350
	(0.05)	(0.05)		
Full time teacher (1 if yes, 0 if no)	0.31	0.35	0.06	0.594
	(0.05)	(0.05)		
Socioeconomic status index (	0.48	0.48	0.00	0.992
1 if below median, 0 if not)	(0.05)	(0.05)		
Cohort				
Second (1 if yes, 0 if no)	0.21	0.25	0.05	0.621
	(0.05)	(0.05)		
Third (1 if yes, 0 if no)	0.37	0.33	-0.05	0.633
	(0.05)	(0.05)		
Observations	84	81		

Standard errors are in parentheses. DIME denotes Diagnostic Instrument for the Measurement of English. CEFR denotes the Common European Framework of Reference. This table shows that randomization was successful at creating two groups that are not different on average. *Differences above 0.25 may denote problems with linearity assumptions.

Table 3: Differences between Teachers who Leave the Sample and those who are Included

	Included	Excluded	Standardized differences in means*	P-value (value for test of equality)
Treatment (1 if treatment, 0 if control)	0.53	0.33	0.29	0.086
	(0.04)	(0.11)		
English level				
DIME total (scale 0 to 120)	64.95	66.73	-0.09	0.589
	(1.16)	(3.41)		
DIME writing (scale 0 to 30)	20.16	21.53	-0.19	0.243
	(0.42)	(1.12)		
DIME speaking (scale 0 to 30)	18.18	19.31	-0.19	0.292
	(0.39)	(0.77)		
DIME listening (scale 0 to 30)	14.04	13.23	0.11	0.457
	(0.37)	(1.29)		
DIME reading (scale 0 to 30)	12.57	12.66	-0.01	0.940
	(0.41)	(1.11)		
CEFR (1 if B1 and 0 if A1+ or A2)	0.42	0.38	0.06	0.713
	(0.04)	(0.11)		
Oral interview score (scale 0 to 10)	4.77	5.00	-0.12	0.478
	(0.11)	(0.28)		
School characteristics				
Teaches in lower secondary school	0.76	0.52	0.35	0.025
(1 if yes 0 if no)	(0.04)	(0.11)		
School size (number of students)	327.03	345.62	-0.05	0.748
	(20.31)	(58.99)		
Individual characteristics				
Gender (1 if male, 0 if female)	0.35	0.35	0.00	0.999
	(0.04)	(0.10)		
Age (in years)	37.95	36.56	0.12	0.478
	(0.70)	(1.71)		
Educ above teacher certification (1 if yes, 0 if no)	0.76	0.80	-0.07	0.693
	(0.03)	(0.09)		
Currently studying (=1 if yes, 0 if not)	0.41	0.40	0.01	0.934
	(0.04)	(0.11)		
Full time teacher (1 if yes, 0 if no)	0.35	0.21	0.24	0.174
	(0.04)	(0.09)		
Socioeconomic status index (1 if below median, 0 if not)	0.47	0.50	-0.03	0.832
	(0.04)	(0.11)		
Cohort				
Second cohort (1 if yes, 0 if no)	0.26	0.05	0.43	0.033
	(0.04)	(0.05)		
Third cohort (1 if yes, 0 if no)	0.33	0.48	-0.20	0.203
	(0.04)	(0.11)		
Observations	144	21		

Standard errors are in parentheses. P-values are tests of the null hypothesis of equality of means. DIME denotes Diagnostic Instrument for the Measurement of English. CEFR denotes the Common European Framework of Reference. This table shows that besides treatment status, the group of teachers that attritted is not statistically different in means from the group of teachers that did not attrit. *Differences above 0.25 may denote problems with linearity assumptions.

fore intervention. Therefore, program effects can be estimated by comparing means between the two groups after program implementation. Estimation is reported with and without controlling for the socioeconomic status (SES) of teachers. Robustness checks are included to rule out that these conclusions are a result of selection bias. Estimation of program effects in the main analysis is robust to the inclusion of the excluded cohort of 17 teachers and to the inclusion of a richer set of controls. Robustness checks including estimation of differences in differences and Manski Lee bounds on applicable data support findings in the main specification.

3.4 Estimating Equation

A comparison of means between treatment and control groups after program implementation estimates the average impact of the program on students. The estimating equation for the program is:

$$y_{ijc} = \mu_c + \beta DT_{ijc} + \gamma X_{jc} + \epsilon_{ijc} \tag{1}$$

Where y is the outcome of interest, i denotes a student, j denotes teacher, c denotes a cohort-level group (cohort and upper vs lower secondary), DT is a dummy which equals one if the teacher was assigned to treatment and zero otherwise, X denotes the socioeconomic status of the teacher and ϵ is an error term. Stratified randomization in the design is exploited and therefore a cohort-level group fixed effect μ_c is introduced. Errors are clustered at the teacher level. For teacher outcomes, the estimating equation is:

$$y_{jc} = \mu_c + \beta DT_{jc} + \gamma X_{jc} + \epsilon_{jc} \tag{2}$$

White robust standard errors are estimated for teacher outcomes. Program ef-

Table 4: Baseline Means on Teacher Characteristics in Evaluation Sample

	Treatment	Control	Standardized differences in means*	P-value** (value for tests of equality)	P-value (SES)***
English proficiency level					
CEFR (1 if B1 and 0 if A1+ or A2)	0.42	0.43	0.02	0.836	0.822
	(0.06)	(0.06)			
Oral interview score (scale 0 to 10)	4.80	4.74	-0.03	0.789	0.982
	(0.16)	(0.17)			
School characteristics					
Teaches in lower secondary school	0.73	0.79	0.11	0.377	0.989
(1 if yes 0 if no)	(0.05)	(0.05)			
School size (number of students)	324.20	330.28	0.02	0.882	0.930
	(28.22)	(29.45)			
Individual characteristics					
Gender (1 if male, 0 if female)	0.33	0.37	0.06	0.586	0.837
	(0.05)	(0.06)			
Age (in years)	37.85	38.06	0.02	0.883	0.780
	(0.87)	(1.12)			
Educ above teacher certification	0.78	0.73	-0.09	0.453	0.498
(1 if yes, 0 if no)	(0.05)	(0.05)			
Full time teacher (1 if yes, 0 if no)	0.32	0.39	0.11	0.338	0.466
	(0.05)	(0.06)			
Marital status (1 if married or lives	0.52	0.56	0.06	0.611	0.679
together, 0 if not)	(0.06)	(0.06)			
Socioeconomic status index	0.47	0.48	0.02	0.876	0.186
(1 if below median, 0 if not)	(0.06)	(0.06)			
Cohort					
Second (1 if yes, 0 if no)	0.23	0.28	0.08	0.498	0.366
	(0.05)	(0.06)			
Third (1 if yes, 0 if no)	0.36	0.30	-0.10	0.412	0.065
	(0.06)	(0.06)			
Observations	77	67			

Standard errors are in parentheses. CEFR denotes the Common European Framework of Reference. This table shows that randomization was successful at creating two groups that are not different on average. *Differences above 0.25 may denote problems with linearity assumptions. **P-values are tests of the null hypothesis of equality of means. ***P-values for tests for the null of hypothesis of equality of means with controls for the socioeconomic status of teachers to consider the possibility of selection on observables due to attrition.

fects are reported with and without controlling for the socioeconomic status of teachers X .

4 The Effects of In-Service Teacher Training on Students' English Proficiency

Table 5: Impact on Student English Proficiency

Dependent variable	Mean on control group	*Impact without controls*			*Impact controlling for teacher SES*		
		Impact	P-value	WMW P-value†	Impact	P-value	WMW P-value†
DIME total	26.91	1.74	0.325	0.000	2.29	0.214	0.000
		(1.77)			(1.84)		
DIME listening	7.85	0.49	0.310	0.014	0.64	0.191	0.000
		(0.48)			(0.49)		
DIME reading	10.54	0.63	0.272	0.010	0.73	0.204	0.000
		(0.58)			(0.58)		
DIME writing	4.82	0.18	0.744	0.166	0.29	0.614	0.015
		(0.54)			(0.58)		
DIME speaking	3.70	0.44	0.457	0.028	0.63	0.306	0.000
		(0.59)			(0.61)		

Tests based on a sample of 718 students (5 per teacher and 2 missing). Standard errors are in parentheses clustered by teacher. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. † WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher ranks than the other group.

The main evaluation question is whether or not the program improves English in students. Table 5 shows estimated program effects for the DIME test. Students of treated teachers score on average 2.3 points more than students in the control group, equivalent to 0.16 standard deviations. The absence of statistical significance may be a result of the lack of power to detect this effect. With the observed intra-cluster correlation of 0.5 and 67 teachers in treatment and 77 teachers in control and a sample of five students per teacher, the evaluation set up has the power to detect a minimum detectable effect of 5 points.

The Wilcoxon-Mann-Whitney ranking test is conducted to test whether the dis-

tributions of treatment and control students differ. It is assumed that the DIME scores can rank students in order of English knowledge. The null of no effects is rejected at the 1 percent level for the total DIME test and the listening, reading, and speaking sections. If the effect size is the point estimate of 2.3 points, then the program provides the additional benefits equivalent to ten weeks of normative progress in English according to the national curriculum in an average period of seven and a half months of exposure.[10] Results are robust to make a simple comparison of treatment and control groups.

5 Inside the Box: Changes in Teacher Subject Knowledge and Classroom Practices

This section explores two underlying mechanisms by which teacher training may affect student learning. Figure 1 shows the logical flow. Teacher training may change what teachers know in two areas: English (subject knowledge) and pedagogy (how to teach). The classroom practices that are explored include class structure and teacher instructional practice. Classroom changes lead to changes in student's use of time, expectations, and private investment, which in turn result in learning.

5.1 Changes in Teacher Subject Knowledge

The average score for control teachers is 63 out of 120 points, which is equivalent to a score of around 506 points on the TOEFL exam and B2 in the Common European Framework of Reference. Table 6 shows program effects before program implementation (baseline), right after training (short run), and at the end of the school year

[10]Normative progress refers to what a student is supposed to learn according to the curriculum. The national curriculum aims at improving the equivalent of 6.66 points the first year and 10 points the second and third years of secondary education. A school year is 40 weeks of class time. Assuming an average of 8.88 points to 40 weeks, then 2.3 points is equivalent to 10 weeks.

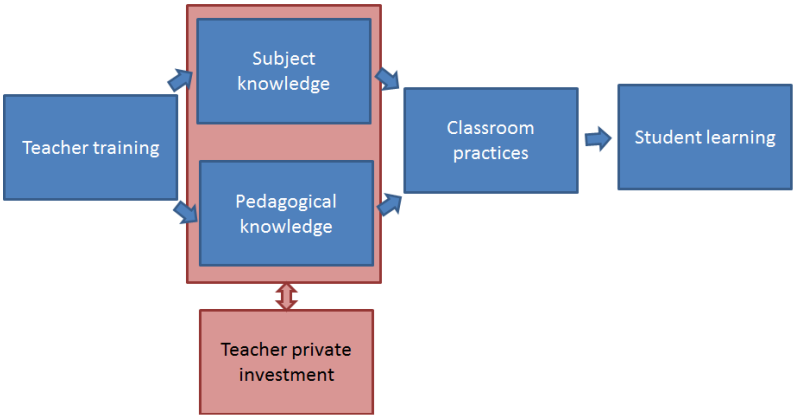

Figure 1: Mechanisms by which Teacher Training Affects Student Learning

(long run). The table includes P-values for the Wilcoxon-Mann-Whitney test. Results are reported both controlling for the SES of teachers and without controls. The baseline results show that there are no differences between the treatment and control groups before the program was implemented. The test for equal distributions is rejected when controlling for teacher SES at baseline. When differences are estimated, consistent effects are found. We take the effects controlling for teacher SES as a caution and find that teachers in the treatment group improved 4.71 points when compared to the control group right after training. This effect comes from improvements in listening and reading. The relative gains in listening and reading were not retained at the end of the school year. Gains in speaking at the end of the school year cannot be rejected. Estimation without controls is consistent with estimation controlling for teachers' SES.

5.2 Changes in Teacher Private Investment

Substitution of program funds among beneficiaries has been documented in education (Das et al., 2013). One policy-relevant hypothesis to test is whether teachers

Table 6: Impact on Teachers' English Proficiency

	Baseline				*Short-run*				*Long-run*			
Dependent variable	Mean on control group	Impact	P-value	WMW P-value†	Mean on control group	Impact	P-value	WMW P-value†	Mean on control group	Impact	P-value	WMW P-value†
Estimates controlling for teacher SES												
Total DIME	63.40	1.91	0.368	0.041	63.17	4.71	0.020	0.000	71.06	1.22	0.697	0.334
		(2.12)				(2.00)				(3.12)		
Listening	13.35	0.93	0.176	0.010	13.24	2.54	0.000	0.000	15.82	-0.74	0.370	0.122
		(0.69)				(0.68)				(0.82)		
Reading	12.35	0.18	0.821	0.626	12.29	1.40	0.045	0.000	14.52	-0.27	0.756	0.364
		(0.77)				(0.69)				(0.86)		
Writing	19.69	0.66	0.421	0.150	19.63	0.67	0.396	0.071	19.54	0.27	0.844	0.440
		(0.81)				(0.79)				(1.36)		
Speaking	18.01	0.15	0.844	0.487	17.99	0.46	0.522	0.170	21.18	1.96	0.183	0.000
		(0.75)				(0.71)				(1.46)		
Estimates without controls												
Total DIME	63.40	2.46	0.254	0.179	63.17	5.48	0.008	0.005	71.06	0.74	0.811	0.527
		(2.14)				(2.03)				(3.12)		
Listening	13.35	1.21	0.091	0.128	13.24	2.77	0.000	0.000	15.82	-0.91	0.271	0.353
		(0.71)				(0.69)				(0.82)		
Reading	12.35	0.32	0.680	0.726	12.29	1.76	0.015	0.032	14.52	-0.37	0.672	0.747
		(0.77)				(0.71)				(0.88)		
Writing	19.69	0.79	0.341	0.413	19.63	0.85	0.291	0.347	19.54	0.27	0.839	0.731
		(0.83)				(0.80)				(1.34)		
Speaking	18.01	0.14	0.857	0.571	17.99	0.46	0.537	0.644	21.18	1.75	0.201	0.047
		(0.80)				(0.74)				(1.37)		

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status. † WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher rankings than those in the other group.

decrease private investment in subject knowledge when they receive training provided by the public education system. If teachers lack incentives to increase subject knowledge, then some government-provided teacher training resources may go to substitute teacher-provided private resources. As a result, program efficiency would decrease. A before-after comparison of English proficiency levels shows that teachers improve their English by 9 points with or without training by the end of the school year. Table 7 shows that teachers without training invested MXN$3,132 in the six months prior to the end of the school year to buy methods or materials to improve their English.[11] Trained teachers invested MXN$1,672 less than non-trained teachers in improving English. Estimations controlling for the socioeconomic status of teachers also show that teachers substitute investment in improving classroom

[11] All investment amounts in the document are expressed in 2013 Mexican pesos

practices from MXN$2,340 to MXN$1,279. This evidence suggests that the private market is more efficient in improving English. This result should be interpreted with caution. The IAPE program implemented in Puebla does not include the follow-up training that the program provides in other states. If the program methodology requires more investment upfront when compared to alternative training programs, then a comparison in the long run may lead to the opposite conclusion.

Table 7: Impact on Teacher Investment

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Private investment (in Mexican pesos)†			
Investment on improving English	3131.80	-1700.72	-1671.73
		(755.91)**	(775.26)**
Investment on improving classroom practices	2339.84	-1053.19	-1060.62
		(655.55)	(642.94)*

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status. † One peso was equivalent to US$.13 as of January 13, 2013 as established by the Mexican Central Bank and published in the *Diario Oficial de la Federación* .

A closely related issue regarding teacher behavior mediating the effects of teacher training on student learning is changes in labor supply. If teachers increase their labor supply and divert resources, this may hinder program effects. We cannot rule out that teachers increased the supply of labor in private schools or private classes. Another concern would be changes in teacher expectations on students which result in changes in private investment. Students who have teachers which expect them to succeed later in life perform better than those that do not (Tosenthal and Jacobson, 1968). No effects on teacher expectations of their students were found. Appendix B shows the results.

5.3 Changes in Classroom Practices

Changes in teacher allocation of time in the classroom are explored for two reasons. First, observing changes in the classroom illuminates student outcomes by determining what teachers changed and how students reacted to these changes. Second, observing classroom changes enables changes to be identified that may lead to an effect on student learning in the long run and explain possible improvements in spoken English. This section includes an analysis of class features in three areas: class structure, instructional practices and its effects on students' behavior in and out of the classroom. Regarding class structure, trained teachers provide students with a more active role in learning by speaking English an additional 14 percent of class time, and by reducing activities that focus on reading and writing 14 percent substituting for activities that focus on listening and conversation. Regarding instructional practices, trained teachers spend 9 percent less time in class monitoring activities that students carry out while seated and increase dynamic activities. Teachers speak English an additional 14 percent of class time. Teachers decrease the class time where students work with textbooks by 7 percent and increase the amount of time spent using didactic materials. Trained teachers show more confidence.

5.3.1 Class Structure

National Center for Education Statistics (2003) video study supports the idea that more time on task, activities where students play an active rather than a passive role, the introduction of new content, and more teacher and student interactions should lead to more learning. Table 8 shows the effects of the program on the structure of the class as measured by the relative weight of the tasks conducted by the teacher. Classes with a trained teacher decrease the percentage of class time spent by the

teacher on answering questions based on the reading from 17 to 9 percent. Teachers decrease the percentage of class time spent leading activities based on students' composing or writing texts from 22 to 13 percent. Trained teachers increase the percentage of class time spent by the teacher on listening and conversation activities from 12 percent to 26 percent. The addition of time on listening and conversation is not statistically different to the reduction on reading, writing and composing (p-value=0.969). Other class structure characteristics including time on task, purpose of the lesson, organization, grading, and pedagogical features are measured to assess changes. There are no effects in these dimensions except for homework revision. 12% of teachers in the control group review homework while only 3% reviews homework in the treatment group (p=0.039 for a test of differences in means). Appendix 7 shows program estimates on these other dimensions measured by the structure of class.

Table 8: Impact on the Structure of the Class

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Role of tasks. Percentage of class time where the teacher...			
Spend on new vocabulary, spelling or completing sentences	11.04	3.83 (3.30)	3.37 (3.16)
Answer questions based on reading	17.10	-8.55 (4.56)*	-8.07 (4.46)*
Spend on student composing/writing texts	21.50	-9.36 (4.24)**	-9.27 (4.36)**
Designate for answer only tasks	8.84	0.68 (3.70)	0.57 (3.79)
Spend on listening/conversation	11.62	13.49 (3.97)***	14.00 (4.15)***

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

5.3.2 Instructional Practices

Tables 9 and 10 show the effects of the program on instructional practices. National Center for Education Statistics (2003) lists the features that enhance student learning. Teachers should devote more time to activities where students are directly involved. Teachers should consider real-life situations and speak English during class (as subject exposure). Teachers should promote student inquiry on the subject. Therefore, more questions and answers in class are associated with better outcomes. The learning climate should be one of respect and where the teacher feels confident. There is no specific recommended allocation of time to use any specific given material in class.

The two most common activities observed in the classes are work while students are seated and lectures. Teachers explicitly instruct students to perform a task in a notebook or books, or to work with other classmates while they are seated at their desks 20 percent of the time and teachers lecture 14 percent of time. Teachers decrease the amount of time that students are sitting at their desks and increase the time spent on dynamic activities by 8 percent of class time.[12] The reduction in student work while seated is not statistically different from the increase in time spent on games and dynamic activities (p-value=0.0993). Teachers spend 7 percent of class time on tasks that consider real-life situations. Teachers in the control group spend 28 percent of the class time speaking English, while teachers in the treatment group spend 43 percent of the time speaking English. In control schools, the most frequently used materials were the blackboard, used 28 percent of the time, followed

[12]Specifically, teachers increase the time spent on three techniques encouraged by the IAPE program: repetition drills, from 8 to 16 percent (p= 0.000); substitution drills, from 1 to 4 percent (p= 0.010); and discrimination drills, from 0.4 to 1.9 percent (p= 0.052). Teachers do not change the amount of time spent on transformation drills (1.4 percent) or personalized drills (3.7 percent).

by work without materials and the textbook, around 16 percent each, notebooks and writing devices, 14 percent, and didactic materials and information and communication technologies, 8 percent. Teachers that participated in the program increased the time that students use didactic materials from 8 to 19 percent and reduced the use of textbooks from 16 to 8 percent. The increase in the use of didactic materials is not statistically different to the reduction in the use of textbooks suggesting a substitution (p-value=0.585). Regarding the learning climate, IAPE teachers showed more confidence, from 8 to 18 percent of the time. No instances of a teacher making an inappropriate remark or ridiculing a particular student were observed.

5.3.3 Effects of Classroom Changes on Student Behavior

Table 11 shows that student time allocation in class is a function of whether students are listening, writing or speaking English, and class climate. Students in the control group spend 32 percent of class time listening to the teacher or another student speaking English. 14 percent of the class time most students are writing English and about 2 percent of the time students are engaged in conversations in English. The program changes the allocation of time to these activities. Students with a trained teacher spend more time listening and conversing, with respective increases of 15 and 4 percentage points. Students spend less time writing decreasing the share of class time in 8 percentage points.

Regarding classroom climate, during 91 percent of class time, most students are paying attention to the teacher but do not show enthusiasm. Training of teachers does not change the amount of time that students show this attitude. Pure enthusiasm is rare. On average, it is observed only in 0.15 percent of class time. Students of treatment teachers increase their expressions of enthusiasm to 0.8 percent of the time. Table 11 shows the estimates of the program's impact on these dimensions.

Table 9: Estimates of Impact of the IAPE Program on Instructional Practices

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Teacher activities. Percentage of class time where the teacher...			
Asks questions	8.93	4.95	5.49
		(2.76)*	(2.98)*
Lectures	14.41	-2.89	-2.71
		(2.40)	(2.34)
Does a demonstration/ presentation of material	5.87	1.49	1.34
		(2.51)	(2.64)
Monitors student work while seated	19.70	-9.06	-8.84
		(3.70)**	(3.68)**
Writes on the blackboard	8.00	-1.19	-1.29
		(1.49)	(1.45)
Reads from a textbook or notebook	8.88	-1.37	-1.05
		(2.34)	(2.26)
Leads role-playing or representations	1.07	2.23	2.00
		(1.64)	(1.64)
Leads games or dynamic activities	3.78	8.14	7.72
		(2.84)***	(2.83)***
Monitors student activities	7.78	-1.66	-1.33
		(2.15)	(2.36)
Context. Percentage of class time where the teacher...			
Considers real-life situations as lesson's content	6.81	0.75	0.38
		(3.02)	(3.02)
Speaks in English	28.23	15.19	14.09
		(3.04)***	(2.94)***

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Table 10: Estimates of the Impact of the IAPE Program on Instructional Practices

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Percentage of class time where the teacher uses for a given activity...			
No materials	16.27	3.65	3.75
		(2.91)	(3.07)
The textbook	15.56	-7.20	-7.48
		(3.36)**	(3.48)**
Notebooks or writing devices	13.91	-2.42	-2.34
		(3.54)	(3.46)
Blackboard	28.08	-4.06	-4.46
		(4.39)	(4.42)
Didactic materials	8.05	10.49	11.27
		(3.64)***	(3.56)***
Information and communication technologies	8.20	2.96	2.87
		(3.73)	(3.91)
Learning climate. Percentage of class time that the teacher...			
Shows confidence, command of class, and stage presence	8.28	9.80	9.37
		(4.27)**	(4.31)**

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Results are robust to the exclusion of SES.

Table 11: Impact on Student Time Allocation in English Class

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Engaged *listening* to the teacher speaking English. or another student	31.60	15.26 (3.42)***	14.52 (3.36)***
Writing English in their notebooks or textbooks	14.48	-7.81 (2.59)***	-7.90 (2.62)***
Engaged in conversations *speaking* English	1.55	4.08 (0.91)***	4.05 (0.94)***
Paying attention to the teacher without showing enthusiasm	91.16	-2.25 (1.84)	-2.49 (1.94)
Showing a clear behavior of happiness or excitement on the learning tasks	0.15	0.69 (0.27)**	0.66 (0.27)**

Tests based on a sample of 718 students (5 per teacher and 2 missing). Standard errors are in parentheses clustered by teacher. Each row represents a different dependent variable in a regression model. Dependent variables are expressed as percentage of class time devoted to a given activity. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Students' expectations and beliefs can lead to changes in investment and therefore learning in the long run. For example, Chiapa, Garrido, and Prina (2012) and Brown, Ortiz-Nunez, and Taylor (2011) both find that higher educational aspirations and career expectations improve students' attainment. The program did not change students' enjoyment of English class or their perception of its utility for their lives. Students with a treatment teacher improved their expectations about the likelihood that they will have a job when they are 30 years old and that they will attend university. The training program does not directly aim to change student expectations. The psychological component of the program, which allows students to believe that they are capable of doing something they thought they could not do, such as speak English, may have influenced expectations. Students with IAPE-trained teachers are more likely to study English on their own and increase the time they spend studying. Appendix D shows program estimates.

6 Robustness Checks

Program effects have been estimated with and without controlling for the SES of teachers. Controls have been introduced to address any potential bias derived from the 13 percent attrition rate observed between the baseline and the follow-up. This section presents four other checks to assess if potentially unobservable characteristics could threaten the validity of the results.

First, we estimate student effects, introducing additional controls and adding the 17 teachers in the omitted cohort in the main analysis. Tables 12 and 13 show estimated program effects. The results are consistent with those in the main analysis. This evidence suggests that attrition does not bias estimates and that observable characteristics are not correlated to treatment or outcomes. This evidence is consistent with identification resulting from exogenous variation in treatment and non-selective attrition.

Table 12: Impact on Student English Proficiency: Adding the 17 Teachers in the Omitted Cohort

Dependent variable	Mean on control group	*Impact without controls*			*Impact controlling for teacher SES*		
		Impact	P-value	WMW P-value†	Impact	P-value	WMW P-value†
Total DIME (score120)	26.96	1.89	0.246	0.000	2.48	0.147	0.000
		(1.62)			(1.70)		
Listening (score 30)	7.91	0.41	0.371	0.027	0.59	0.201	0.000
		(0.45)			(0.46)		
Reading (score 30)	10.60	0.61	0.248	0.009	0.73	0.175	0.000
		(0.53)			(0.53)		
Writing (score 30)	4.83	0.29	0.558	0.067	0.42	0.441	0.001
		(0.50)			(0.54)		
Speaking (score 30)	3.61	0.57	0.297	0.004	0.74	0.199	0.000
		(0.55)			(0.57)		

Tests based on a sample of 788 students (144 teachers, adding 17 teachers in the omitted cohort, of which only 14 have follow up information, resulting in 158 teachers. Five student per teacher and two missing). Standard errors are in parentheses clustered by teacher. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. † WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher ranks than the other group.

Table 13: Impact on Student English Proficiency: Introducing Additional Controls

Dependent variable	Mean on control group	*Impact without controls*			*Impact controlling for teacher SES*		
		Impact	P-value	WMW P-value†	Impact	P-value	WMW P-value†
Total DIME (score120)	26.91	1.75	0.261	0.000	2.24	0.167	0.000
		(1.55)			(1.62)		
Listening (score 30)	7.85	0.48	0.278	0.014	0.63	0.163	0.000
		(0.44)			(0.45)		
Reading (score 30)	10.54	0.64	0.236	0.010	0.73	0.176	0.000
		(0.54)			(0.54)		
Writing (score 30)	4.82	0.18	0.729	0.166	0.28	0.607	0.003
		(0.51)			(0.55)		
Speaking (score 30)	3.70	0.45	0.399	0.028	0.61	0.271	0.000
		(0.53)			(0.55)		

Tests based on a sample of 718 students (5 per teacher and 2 missing), controlling for student's age, gender, years of education, has oportunidades scholarship or not, mother's education in years, and if school is located in home locality. Standard errors are in parentheses clustered by teacher. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. † WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher ranks than the other group.

Second, for the short-run results in terms of English proficiency of the teachers, there is information for all but two of the 165 teachers present at baseline. If the inclusion of these teachers result in estimates that differ from those from the restricted set of 144 teachers after attrition, then bias may be present in the main specification. Table 14 shows the estimation of effects based on 163 out of 165 teachers. Estimates including all observations are not statistically different from those estimated on the restricted sample of 144 teachers. Therefore attrition bias in the main specification is unlikely.

Third, program effects are calculated using difference in differences for the indicators for which there is available baseline data, which is teacher DIME and questionnaire but excludes student outcomes or classroom observations. Difference in differences allows for program estimation, controlling for characteristics that are invariant in time. Differentiating equation 2 with its lag:

Table 14: Impact of the IAPE Program on Teachers' English Proficiency

Dependent variable	Mean on control group	Impact	P-value	WMW P-value†
Total DIME	63.39	5.67	0.004	0.004
		(1.97)		
Listening	12.93	3.22	0.000	0.000
		(0.66)		
Reading	12.24	1.93	0.005	0.021
		(0.68)		
Writing	19.98	0.46	0.540	0.666
		(0.75)		
Speaking	18.23	0.38	0.570	0.839
		(0.67)		

Estimation based on 163 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model without controls. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. † WMW P-values denotes correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that one group have larger values on the DIME test than the other.

$$\Delta y_{jc} = \Delta\mu_c + \beta DT_{jc} + \Delta X_{jc} + \Delta\epsilon_{jc} \tag{3}$$

Where: $\Delta y_{jc} = y_{jc1} - y_{jc0}$ is the value added to teacher subject knowledge, $\Delta\mu_c = \mu_{c1} - \mu_{c0}$ and $\Delta\epsilon_{jc} = \epsilon_{jc1} - \epsilon_{jc0}$. Note that $DT_{jc1} - DT_{jc0} = DT_{jc1} - 0 = DT_{jc1}$ denoted as DT_{jc} and $\Delta X_{jc} = X_{jc1} - X_{jc0} = 0$.

Time period 1 denotes the period after program intervention, at the end of the school year. Time period 0 denotes baseline observation. Simplifying notation making $\nu_c = \Delta\mu_c$ and $u_{jc} = \Delta\epsilon_{jc}$ equation 3 can be written as:

$$\Delta y_{jc} = \nu_c + \beta DT_{jc} + u_{jc} \tag{4}$$

The first difference estimation is equivalent to the estimation of program effects using teacher fixed effects because the panel has two periods. Tables 15 and 16 show the estimates for teachers' English proficiency and investment. The results are

consistent with the main specification both when these are significant and when they are not. Therefore, estimates on gains in teacher English proficiency and investment are likely to hold even in the presence of systematic attrition related to time-invariant observable and unobservable characteristics. There is no baseline information for student outcomes or classroom observations and this check is not possible.

Table 15: Impact on Changes in Teachers' English Proficiency

	Short-run			*Long-run*		
Dependent variable	Impact	P-value	WMW P-value†	Impact	P-value	WMW P-value†
Total DIME	3.03	0.001	0.000	-2.27	0.465	0.601
	(0.92)			(3.10)		
Listening	1.57	0.000	0.000	-2.22	0.020	0.065
	(0.40)			(0.94)		
Reading	1.44	0.000	0.000	-0.88	0.383	0.630
	(0.40)			(1.01)		
Writing	0.06	0.893	0.347	-0.58	0.673	0.796
	(0.46)			(1.38)		
Speaking	0.31	0.400	0.043	1.41	0.325	0.105
	(0.37)			(1.43)		

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a difference-in-differences regression model controlling for the socioeconomic status of teachers. Columns (2) to (4) include estimates in the short run (right after training is over). Columns (5) to (7) include estimates in the long run (end of school year) *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status. † WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher ranks than the other group.

Fourth, Manski-Lee bounds as described by Lee (2009) are calculated. The intuition behind Manski-Lee bounds is that within each stratum, the sample for the treatment group is trimmed to make attrition rates equal among both treatment and control groups. Trimming teachers with the highest DIME scores in the treatment group provides a lower bound.[13] Lee bounds can be tightened by introducing controls in estimation that have explanatory power on attrition. Therefore, strata fixed effects are included to tighten bounds (Tauchmann, 2012). Estimation results in

[13]Estimation is performed with the command lee-bounds in STATA.

Table 16: Impact on Changes in Teacher Investment

Dependent variable	Impact	P-value	WMW P-value[†]
Private investment			
Buying methods or materials to improve your English	-861.11	0.154	0.279
(In mexican pesos[‡])	(601.44)		

Tests Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a difference-in-differences regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. [†] WMW P-values correspond to Wilcoxon-Mann-Whitney tests of the null hypothesis that students in one group have higher ranks than the other group. [‡] One peso was equivalent to US$.13 as of January 13, 2013 as established by the Mexican Central Bank and published in the *Diario Oficial de la Federación* .

Table 17: Manski-Lee Bounds

	Lower Coefficient	Upper Coefficient	Lower Bound	Upper Bound
Percentage of class time that	10.97	19.25	4.50	25.34
the teacher speaks English	(3.93)***	(3.70)***		
Investment on improving English	-2451.84	-1523.23	-3844.00	-94.25
(in mexican pesos[†])	(831.71)***	(853.71)**		

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents upper and lower bound estimation of Manski-Lee bounds for a different dependent variable. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. [†] One peso was equivalent to US$.13 as of January 13, 2013 as established by the Mexican Central Bank and published in the *Diario Oficial de la Federación* . Estimates for other dependent variables not shown because bounds are too wide to be informative.

wide intervals for most estimates on program effects. In spite of this limitation, intervals of estimates of program effects on the amount of time that teachers speak English in class and on changes on investment confirm previous findings. Table 17 shows estimates for these two outcomes. For the rest, intervals are too wide to be informative and are not reported. Lee bounds are not calculated for students, as that would require further assumptions on what represents a teacher with better student results.[14]

Teacher results are robust even in scenarios where selective attrition is not addressed with observable characteristics. These robustness checks cannot be carried out for students because there are no baseline data available. Student outcomes are very likely correlated with teacher outcomes, and patterns are likely to be similar. No program effects are calculated with propensity score matching because balance at baseline on observables implies that results will not change matching for observable characteristics.

7 Summary, Discussion, and Policy Implications

The Inter-American Partnership for Education Program in Mexico aims to improve English proficiency and change pedagogical practices of English teachers in public schools. The goal of the program is to provide students with the English skills to be competitive in a more globalized labor market and benefit from increased access to resources available in English, such as publications, instructions for technology adoption, and communication. Program effects are estimated through a randomized controlled trial. This study shows that teacher training led to improvements in

[14]The O'Brien (1984) test for multiple outcomes is not carried out because it requires the assumption of effects in a specific direction. For a large number of outcomes, especially those related to time allocation in class, there are no theoretical priors to encourage one-sided tests.

teacher knowledge and in pedagogical techniques. Classroom activities changed by the end of the school year, with more time spent on games or dynamic activities and less time on tasks in their notebooks or books while sitting at their desks. Trained teachers spend more class time speaking English. As a result, students improved their English by around 0.16 standard deviations.

These findings support those in the education literature. Teacher training may lead to improvements in student learning but only under certain conditions. Teacher training is likely to need aligned teacher incentives and continuous training to achieve sustained gains. The decrease in teacher private investment to improve English may be a result of a lack of incentives. Teachers improved their reading and listening skills in English, but the control group caught up with them by the end of the school year. Improving the quality of education provision is challenging, and many studies show that complementarities play a key role in human capital formation.

The long-run gains of the program will depend on how sustainable the benefits are. To get a sense of the sensitivity of the analysis to assumptions, assume the effects of the program increase in English proficiency of 2.3 out of 120 points, an unemployment rate of 8.5 percent, that half of the students for each teacher benefits from a return to English of 24 percent constant over time, and that students earn a constant MXN$4,000 per month, which is the urban salary for high school graduates.[15] If the teacher only benefits one cohort of 4 groups with 33 students and the students obtain returns to English for five years, then the return rate is 196 percent. If the program provides benefits to students less than 24 months and teachers forget

[15]Salary according to Panorama Educativo, INEE (Institute of the Evaluation of Education, Mexico). Unemployment rate according to INEGI. Encuesta Nacional de Ocupación y Empleo, second quarter of 2013. The rate of return to skill acquisition according to Azam, Chin, and Prakash (2013) is 14 percent for speaking little English and 34 percent for speaking fluent English. A midpoint is chosen.

within one school year and do not benefit any other generations, then the investment is a loss. This calculation shows just how sensitive the calculation of benefits is to assumptions. If students continue to learn and benefits are larger and sustained over time, the benefits will be much larger. On the other hand, if teachers and students forget, then the benefits of the program may be null. Therefore attention should be given to complements that lead to sustainable learning.

Many areas remain to be explored. It is important to explore whether changes in teacher and student behavior are sustainable over time and, if so, for how long. Another area to explore is how complementarities affect the effectiveness of teacher training. It is especially relevant to explore how to link teacher incentives to training. Answering these questions may provide clues on how to provide equitable access to skills acquisition for all students.

References

Angrist, J.D., and V. Lavy. 2001. "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools." *Journal of Labor Economics* 19:343–69.

Azam, M., A. Chin, and N. Prakash. 2013. "The Returns to English-Language Skills in India." *Economic Development and Cultural Change* 61:335 – 367.

Bleakley, H., and A. Chin. 2004. "Language Skills and Earnings: Evidence from Childhood Immigrants." *The Review of Economics and Statistics* 86:pp. 481–496.

Brown, S., A. Ortiz-Nunez, and K. Taylor. 2011. "What will I be when I grow up? An analysis of childhood expectations and career outcomes." *Economics of Education Review* 30:493 – 506.

Carnoy, M., et al. 2008. "How schools and students respond to school improvement programs: The case of Brazil's PDE." *Economics of Education Review* 27:22 – 38.

Chiapa, C., J.L. Garrido, and S. Prina. 2012. "The effect of social programs and exposure to professionals on the educational aspirations of the poor." *Economics of Education Review* 31:778 – 798.

Chingos, M.M., and P.E. Peterson. 2011. "It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness." *Economics of Education Review* 30:449 – 465.

Cummins, J. 1986. "Empowering minority students: A framework for intervention." *Harvard Educational Review* 56:18–36.

Darling-Hammond, L., et al. 2009. "Professional Learning in the Learning Profession: A Status Report on Teacher Development in the United States and Abroad." Working paper, National Staff Development Council and The School Redesign Network at Stanford University.

Das, J., et al. 2013. "School Inputs, Household Substitution, and Test Scores." *American Economic Journal: Applied Economics* 5:29–57.

de Mel, S., D. McKenzie, and C. Woodruff. 2014. "Business training and female enterprise start-up, growth, and dynamics: Experimental evidence from Sri Lanka." *Journal of Development Economics* 106:199 – 210.

Ellis, R. 1986. *Second Language Acquisition. Oxford Introductions to Language Study*. ISBN 978-0-19-437212-1, Oxford, New York: Oxford University Press.

Galdo, J., and A. Chong. 2012. "Does the quality of public-sponsored training programs matter? Evidence from bidding processes data." *Labour Economics* 19:970 – 986.

Galliher, R., et al. 1995. "Preparing Technical Educators for Interactive Instructional Technolgies: A Review of Research and Practice." Working paper.

Garet, M., et al. 2011. "Middle school mathematics professional development impact study: Findings after the second year of implementation." Working paper, Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Goldhaber, D., S. Liddle, and R. Theobald. 2013. "The gateway to the profession: Assessing teacher preparation programs based on student achievement." *Economics of Education Review* 34:29 – 44.

Goldschmidt, P., and G. Phelps. 2010a. "Does teacher professional development affect content and pedagogical knowledge: How much and for how long?" *Economics of Education Review* 29:432 – 439.

—. 2010b. "Does teacher professional development affect content and pedagogical knowledge: How much and for how long?" *Economics of Education Review* 29:432 – 439.

Gonzalez, L. 2005. "Nonparametric Bounds on the Returns to Language Skills." *Journal of Applied Econometrics* 20:pp. 771–795.

Hamalainen, K., R. Uusitalo, and J. Vuori. 2008. "Varying biases in matching estimates: Evidence from two randomised job search training experiments." *Labour*

Economics 15:604 – 618, European Association of Labour Economists 19th annual conference/Firms and Employees.

Harris, D.N., and T.R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95:798–812.

He, F., L. Linden, and M. MacLeod. 2008. "How to Teach English in India: Testing the Relative Productivity of Instruction Methods within the Pratham English Language Education Program." *Colubia University Working Paper*, pp. .

Hotz, V.J., G.W. Imbens, and J.H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125:241 – 270, Experimental and non-experimental evaluation of economic policy and models.

Hsiao, C., Y. Shen, B. Wang, and G. Weeks. 2008. "Evaluating the effectiveness of Washington state repeated job search services on the employment rate of prime-age female welfare recipients." *Journal of Econometrics* 145:98 – 108, The use of econometrics in informing public policy makers.

Imbens, G.W., and J.M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47:5–86.

INEGI. 2012. "El Ingreso y el Gasto Público en México 2012." Working paper No. 24.

Jacob, B.A., and L. Lefgren. 2004. "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago." *Journal of Human Resources* 39.

Lee, D.S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *The Review of Economic Studies* 76:pp. 1071–1102.

Mano, Y., J. Akoten, Y. Yoshino, and T. Sonobe. 2013. "Teaching KAIZEN to Small Business Owners: An Experiment in a Metalworking Cluster in Nairobi.", May, pp. .

Mason, G., B. O'Leary, and M. Vecchi. 2012. "Certified and uncertified skills and productivity growth performance: Cross-country evidence at industry level." *Labour Economics* 19:351 – 360.

Metzler, J., and L. Woessmann. 2010. "The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation." IZA Discussion Papers No. 4999, Institute for the Study of Labor (IZA).

Munshi, K., and M. Rosenzweig. 2006. "Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy." *American Economic Review* 96:1225–1252.

Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *The Economic Journal* 120:F187–F203.

National Center for Education Statistics. 2003. "Teaching Mathematics in seven Contries. Results from the TIMSS 1999 Video Study." Working paper No. NCES 2003-013, U. S. Department of Education. Institute of Education Sciences., March.

Nielsen, S. 2010. "Vocational Education and Training Teacher Training." In P. Peterson, E. Baker, and B. McGaw, eds. *International Encyclopedia of Education (Third Edition)*. Oxford: Elsevier, third edition ed., pp. 503 – 512.

O'Brien, P.C. 1984. "Procedures for comparing samples with multiple endpoints." *Biometrics* 40:pp. 1079–1087.

Omaggio, A. 1986. *Teaching language in context: Proficiency oriented instruction*. Boston, MA: Heinle and Heinle Publishers.

Presidencia de la República. 2013. Working paper, Presidencia de la República, Estados Unidos Méxicanos.

Rockoff, J.E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *The American Economic Review* 94:pp. 247–252.

Stallings, J. 1980. "Allocated Academic Learning Time Revisited, or beyond Time on Task." *Educational Researcher* 9:pp. 11–16.

Sunardi, O., M. Widyarini, and J.H. Tjakraatmadja. 2012. "The Impact of Sales Forces Training Program to Employees Behaviour Styles (A Quasi-experimental Case Study In a Medium Sized Enterprise)." *Procedia Economics and Finance* 4:264 – 273, International Conference on Small and Medium Enterprises Development with a Theme? Innovation and Sustainability in SME Development? (ICSMED 2012).

Suzuki, A., H.N. Vu, and T. Sonobe. 2013. "Willingness to pay for managerial training: A case from the knitwear industry in Northern Vietnam." *Journal of Comparative Economics*, pp. .

Tauchmann, H. 2012. "LEEBOUNDS: Stata module for estimating Lee (2009) treatment effect bounds." Statistical Software Components, Boston College Department of Economics.

Tomlinson, P., A. Hobson, and A. Malderez. 2010. "Mentoring in Teacher Education." In E. in Chief: Penelope Peterson, E. Baker, and B. McGaw, eds. *Interna-*

tional Encyclopedia of Education (Third Edition). Oxford: Elsevier, third edition ed., pp. 749 – 756.

Tosenthal, R., and L.F. Jacobson. 1968. "Teacher Expectations for the Disadvantaged." *Scientific American* 218.

Vegas, E., and J. Petrow. 2008. *Raising Student Learning in Latin America: The Challenge for the Twenty-First Century*. The World Bank.

Wayne, A.J., et al. 2008. "Experimenting With Teacher Professional Development: Motives and Methods." *Educational Researcher* 37:469–479.

Appendix A Characteristics of the IAPE Teacher Training Program

Darling-Hammond et al. (2009) review the literature and conclude that the minimal conditions for a teacher training program to be effective are: (i) it should be intensive enough to cause a change in teacher behavior (over 50 hours), (ii) it should be connected to practice, (iii) it should be continuous, and (iv) it must be aligned with teacher incentives. The program is described in the context of these four conditions.

First, the IAPE program provides 100 hours of intensive training, 80 of which are devoted to intensive English instruction and 20 to pedagogical training. Although the program may not be intensive enough to make teachers completely fluent in English, the amount of time is enough to make it feasible to observe improvements. We find no evidence on how many hours a Spanish-speaking learner would need to be exposed to English to master it. Omaggio (1986) estimates that a student of average aptitude in the United States will require about 480 hours to achieve an advanced level of proficiency in Spanish. Cummins (1986) states that when exposed to a foreign language for more hours per day on a consistent basis, the learning process is different and often more efficient.[16]. More time exposed to the language and more frequency of exposure improve learning outcomes (Galliher et al., 1995; Omaggio, 1986).

Second, the program has six characteristics that aim to change classroom practices. The first characteristic is that teachers pledge to speak only English during their course of study and teachers are encouraged to only use English as the language of instruction in their own practices. The second characteristic is that the

[16]For a discussion on the cognitive approach and brain activity related to language acquisition please see Ellis (1986)

program demonstrates and encourages the use of dynamic activities sustained in repeated interactions under a rhythm. For example, the teacher states "I am feeling happy", then student 1 says "I am feeling energetic", student 2 says "I am feeling relaxed", and so forth. Proponents of the method claim that using these types of strategies can increase class participation from an average 4 to 5 times per student to up to 60 times per class under optimal conditions. The third characteristic is that the program encourages teachers to show enthusiasm when providing feedback to students, to look confused when they are wrong, to correct them and make them repeat. The fourth characteristic is that the program encourages humor and drama in class and encourages the teacher to be active while delivering the lesson in order to keep students engaged and speaking. The fifth characteristic is that the system introduces the philosophy that, in the classroom, the student in not an outside observer, but an active participant. Increased participation and unpredictability together make the discipline of learning a language real. The program encourages changes in classroom culture by maintaining an active choreography that helps remove traditional barriers between teachers and students. The sixth characteristic is that the program and techniques encourage teachers and students to rid themselves of the inhibitions often associated with language learning and simply speak. By removing inhibitions, students participate more and therefore increase production.[17]. These program features are relevant as there is evidence that the contents of a program are likely to strongly influence program effectiveness. Galdo and Chong (2012) find that there is a strong correlation between training expenditures per trainee and labor market outcomes and Goldhaber, Liddle, and Theobald (2013) and Nielsen (2010) finds that the quality of pre-service training programs explain differences in

[17]For more information on the Rassias method on which the program is based, please go to http://rassias.dartmouth.edu/method/

teacher effectiveness.

Third, the program is designed to offer periodic follow ups. The evaluation team decided that training for mentors should be postponed as contamination into control teachers would had been very likely. As a result, the program was incorporated without the follow-up components for the school year studied. IAPE procedure is that, after training, participants in the IAPE Intensive English course agree to actively engage in a virtual community where they provide support to each other and commit to continue to teach English in public schools for another three years. The program also contemplates the supply of post-training workshops with the frequency and intensity determined by the state, combining the experiences of both the IAPE Teachers' Collaborative and IAPE Intensive alumni. Evidence in the past shows that the lack of continuous and proper teacher support may lead to losses in investment. For example Goldschmidt and Phelps (2010b) find that teacher training in the United States leads to significant knowledge growth between pre- and post-assessments but that knowledge retention was hindered six months later. For a review on the conditions for productive mentoring programs, see Tomlinson, Hobson, and Malderez (2010).

Fourth, the program is currently not aligned with teacher incentives. The main path through which the government provides incentives for teachers to receive in-service training is the Carrera Magisterial (CM) program, which gives a weight of up to 20 percent to teacher training. The program allows for promotion every 3 to 4 years and provides salary increases of between 11 and 21 percent. The program validates training programs, certificates and completion of degrees as long as they are listed in a catalog of programs previously revised and approved by the Ministry of Education and the teachers union. IAPE is currently not listed in the catalog.

The second path through which teachers receive incentives for in-service training is the Escalafon del Magisterio (EM) law. The EM law states the rules to promote teachers either to principals or to make them able to take posts in the state capitals. The law establishes a point system where teacher education is given a weight of 45 percent of the score. Teacher education is measured as program completion of official programs, with more points given for tracks leading to degrees with a focus on pedagogical content. The absence of teacher incentives to participate in the program was evident on program applications, as about one-third of qualifying teachers applied to the program. Communication problems between the state and the teachers cannot be ruled out, but teachers received an email invitation and an invitation from the regional coordinator. The team also worked with the teachers union to ensure that all potential communication mechanisms with the teachers were exploited. Many other rounds of invitations were sent to the teachers but interest never exceeded the number of places available.

Incentives for training affect not only participation, but also effectiveness. For example, Muralidharan and Sundararaman (2010) study the effects of teacher incentives on student performance in India and find that teacher incentives were especially effective among teachers with more training but that training alone was not a predictor of student improvement.

The role of incentives on the effects of training is not exclusive to in-service teachers. Suzuki, Vu, and Sonobe (2013) finds that individuals with high prior demand for training to increase productivity in knitwear in rural Vietnam benefited more. de Mel, McKenzie, and Woodruff (2014) find that a training program for women to start their own businesses caused changes in business practices only in those that received a grant to own the business.

Appendix B Teacher Labor Supply and Expectations

This section shows estimates of the effects of teacher training on teacher labor supply and expectations. Table 18 shows that random allocation of teacher to treatment and control groups created two groups that are not different before treatment teachers.

Table 18: Baseline Means on Teacher Labor Supply and Expectations

	Treatment	Control	Standardized differences in means*	P-value** (value for tests of equality)	P-value (SES)***
Labor supply: Do you currently ...(1 if yes, 0 if not)					
Teach in other public schools?	0.15	0.16	0.02	0.868	0.853
	(0.04)	(0.04)			
Teach in other private schools/private	0.03	0.17	0.37	0.002	0.005
English classes	(0.02)	(0.04)			
Make translations or other English classes	0.19	0.14	-0.10	0.397	0.216
English classes	(0.04)	(0.04)			
Expectations: How likely is it that most of your students...(1 if very likely or likely, 0 if little likely or not likely)					
Will have a job when they are 30 ?	0.62	0.66	0.06	0.609	0.459
years old	(0.05)	(0.05)			
Will graduate from high school?	0.70	0.78	0.13	0.265	0.219
	(0.05)	(0.05)			
Attend university?	0.30	0.26	-0.07	0.571	0.613
	(0.05)	(0.05)			
Observations	77	67			

Standard errors are in parentheses. This table shows that the means of impact indicators between treatment and control groups were not significantly different before program implementation. *Differences above 0.25 may denote problems with linearity assumptions. **P-values are tests of the null hypothesis of equality of means.***P-values for tests for the null of hypothesis of equality of means with controls for the socioeconomic status of teachers to consider the possibility of selection on observables due to attrition.

Teachers may change the number of hours worked. The cost of an increased workload could offset the benefits to the students. One possibility is that teachers change the time allotted to teaching in other schools or other non-teaching jobs. Fifteen percent of the teachers in the control group teach in other public schools, 7 percent teach in private schools or provide private English classes, and 3 percent of teachers do translations or other non-teaching work. 10 percent of teachers that participated in the IAPE program did translations or other non-teaching work. This

difference does not hold if socioeconomic status is removed as a control.

Table 19: Impact on Teacher Investment and Labor Supply

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Labor supply. Did you have any income in the last month derived from ...(1 if yes, 0 if no)			
Teaching in other public schools	0.15	0.07	0.06
		(0.07)	(0.07)
Teaching English in other private schools	0.07	0.01	0.01
/private English classes		(0.04)	(0.04)
Made translations or other	0.03	0.07	0.07
non-teaching work		(0.04)	(0.04)*

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represent a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status. † One peso was equivalent to US$.13 as of January 13, 2013 as established by the Mexican Central Bank and published in the *Diario Oficial de la Federación* .

Students who have teachers who expect them to succeed later in life perform better than those who do not (Tosenthal and Jacobson, 1968). Table 20 shows that 75 percent of the teachers believe it is likely that their students will have a job when they are 30 years old and 32 percent believe that it is likely their students will attend a university. There are no effects of the program on teacher expectations of students at the end of the school year.

Table 21 shows program effects calculated with difference-in-differences. Difference-in-differences allow for program estimation controlling for characteristics that are invariant in time. Results show that there were no changes in the labor supply or teacher expectations except for the probability that a teacher increases labor supply in other private schools or by teaching private classes. At baseline, 17 percent of teachers in the control group were teaching in other private schools/private English classes, and this fell to 7 percent in the follow-up. In the treatment group, however, teachers teaching in other private schools/private English classes increased from 3

Table 20: Impact on Teacher Expectations of Students

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
How likely is it that most of your students...(1 if very likely or likely, 0 if little likely or not likely.)			
Will have a job when they are 30 years old?	0.75	-0.01	-0.01
		(0.08)	(0.08)
Attend university?	0.32	-0.00	-0.00
		(0.08)	(0.08)

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

to 8 percent. As a result, the difference in levels is not detected, but once initial differences are incorporated, the change becomes clear. It cannot be ruled out that the program caused increases in teacher labor supply in private schools or private classes.

Appendix C Impact of the Program on Class Structure

This section describes dimensions measured related to the structure of class where no program effects were found. Table 22 shows that the program did not have an effect on the time on task or the purpose of the lesson. Time on task is 95 percent of class time. Teachers spend less than 1 percent of time outside the classroom without a teaching purpose. Regarding the purpose of the class, teachers spend 32 percent of their time practicing a previously covered topic but only 23 percent practicing new content. The time presenting new content is about the same as reviewing content that has been covered in previous classes - 18 percent. There is no evidence of program effects on the purpose of the class.

Table 21: Difference-in-differences Impact on Changes in Labor Supply and Expectations

Dependent variable	Impact
Labor supply. Did you have any income in the last month derived from...(1 if yes, 0 if no)	
Teaching in other public schools	0.07
	(0.06)
Teaching in other private schools/	0.16
private English classes	(0.05)**
Making translations or other non-teaching work	0.03
	(0.07)
Teacher expectations	
Will have a job when they are 30 years old?	0.02
	(0.09)
Attend university?	-0.04
	(0.09)

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a differences in differences regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. ‡ One peso was equivalent to US$.13 as of January 13, 2013 as established by the Mexican Central Bank and published in the *Diario Oficial de la Federación* .

Table 23 shows the effects of teacher training on organization, grading, homework, and other pedagogical features. Regarding class organization, 55 percent of the time, teachers address the whole class. The time spent addressing an individual student rose from 17 to 23 percent. The time the teacher addresses a group remains at 1 percent in the two groups. There are no program effects on how much the teacher encourages students to correct each other.

Teachers grade according to the following weights: 37 percent for exams, 29 percent for homework, 14 percent for participation, 10 percent for presentation, and 9 percent for attendance and others. The program does not affect the way that teachers grade.

The program estimated effects show that trained teachers assign 3 percentage points more to participation but only when controlling for the socioeconomic status of the teacher. Regarding pedagogical features, it is found that 47 percent of teachers

in the control group assign homework and 12 percent reviews homework during class. There are no program effects on homework-related activities except for the share of teachers that review homework. The program reduces homework revision from 12 percent to 3 percent. Forty seven percent of teachers state a goal or purpose of the lesson, and 16 percent present a summary of the key aspects of the lesson. The program does not seem to have an effect on any of these pedagogical features.

Table 22: Impact on the Structure of Class

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Time on task			
Length of the lesson (In minutes)	41.13	-1.36	-1.20
		(1.90)	(1.95)
Time on task (Percentage of class time)	94.88	-0.48	-0.74
		(1.38)	(1.47)
Time the teacher spends outside the classroom without a teaching purpose (Percentage if class time)	0.81	-0.14	-0.15
		(0.72)	(0.78)
Purpose of the lesson. Percentage of class time where the teacher			
Reviews previously presented content (in a previous lesson)	17.40	-2.66	-2.42
		(4.32)	(4.46)
Presents new content	18.09	-5.73	-5.73
		(4.21)	(4.32)
Practices a previously covered topic	31.74	1.45	-1.79
		(6.34)	(6.26)
Practices new content	22.72	3.79	5.93
		(5.98)	(6.00)
Spends on improving English grammar	14.06	6.45	6.15
		(4.56)	(4.32)

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Table 23: Impact on the Structure of Class

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Organization. Percentage of class time where the teacher...			
Addresses the whole class	55.12	-0.81	-1.15
		(3.82)	(3.92)
Addresses a group of students	1.30	0.77	0.54
		(0.92)	(0.91)
Addresses an individual student	17.22	4.91	5.60
		(3.21)	(3.28)*
Encourages students to correct each other	0.62	0.36	0.19
(1 if any, 0 if nothing)		(0.37)	(0.33)
Percentage of grading weight given to...			
Exams	37.40	-3.03	-3.14
		(2.53)	(2.58)
Homework	29.41	0.48	-0.27
		(2.22)	(2.22)
Participation	14.27	2.37	2.78
		(1.58)	(1.59)*
Presentations	9.74	-0.70	-0.76
		(1.38)	(1.42)
Attendance	9.18	0.89	1.40
		(1.68)	(1.80)
Pedagogical features. During class, the teacher...(1 if yes, 0 if no)			
Assigns homework.	0.47	-0.13	-0.12
		(0.08)	(0.09)
Reviews homework.	0.12	-0.09	-0.10
		(0.05)*	(0.05)*
States a goal/purpose of the lesson	0.47	-0.12	-0.11
		(0.09)	(0.09)
Presents a summary of the key aspects of	0.16	-0.05	-0.06
the lesson		(0.06)	(0.06)

Tests based on a sample of 144 teachers. Robust standard errors are in parentheses. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Appendix D Student Investment, Expectations, and Beliefs

Student investment

If the program successfully motivates students to learn English, students may change the amount of time and resources they devote to improving English outside the classroom. Increases in student investment in the short run may result in benefits in the long run. Table 24 shows estimations of program effects. Students with IAPE-trained teachers are more likely to study English on their own and increase the time they study. There is no evidence of an effect on the likelihood of students taking English classes outside of school or the time they spend on homework.

Table 24: Impact on Student Investment

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
How many hours did you spend last week ...			
Studying English on your own	0.40	0.26	0.28
		(0.13)*	(0.13)**
Having a conversation in English	0.37	0.05	0.08
		(0.18)	(0.19)
Doing homework	2.09	0.05	0.04
		(0.20)	(0.20)
Do you currently...(1 if yes, 0 if no)			
Study English on your own?	0.19	0.08	0.10
		(0.04)*	(0.04)**
Take English classes out of school?	0.07	0.01	0.01
		(0.02)	(0.02)

Tests based on a sample of 718 students (5 per teacher and 2 missing). Standard errors in parenthesis clustered by teacher. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 level, respectively. SES denotes socioeconomic status.

Student Expectations and Beliefs

The program may change student expectations and beliefs, which in turn can affect private investment. If the program makes a student believe that English class is useful in life or increases her expectations about future use of the language, then the student may study longer hours or put more attention in class leading to enhanced learning. Likewise, if students find they enjoy English class more, they may choose to devote more time related to the activity. Expectations of attending university increase. This result is not robust to exclude socioeconomic status as a control. The program did not change students' perceptions of how much they like English class or how useful it is for their lives. Students with an IAPE teacher change their expectations about the likelihood that they will have a job when they are 30 years old and that they will attend university. The IAPE program does not directly aim to change students' expectations. The psychological component of the program, which allows students to believe that they are capable of doing something they thought they could not do, such as speak English, could have influenced expectations. Table 25 show estimates of program effects on these dimensions.

Table 25: Impact on Student Expectations and Beliefs

Dependent variable	Mean on control group	Impact without controls	Impact controlling for teacher SES
Beliefs. If 10 is a lot and 0 is nothing, how much do you...			
Like English class?	7.76	0.16	0.16
		(0.20)	(0.20)
Think English is useful in life?	8.99	0.06	0.08
		(0.13)	(0.13)
Expectations. How likely is it that you...(1 if very likely or likely, 0 if little likely or not likely)			
Will have a job when you are 30 years old?	0.82	0.09	0.09
		(0.03)***	(0.03)***
Will attend university?	0.63	0.07	0.09
		(0.05)	(0.05)*

Tests based on a sample of 718 students (5 per teacher and 2 missing). Standard errors are in parentheses clustered by teacher. Each row represents a different dependent variable in a regression model. *, **, *** indicate that the estimates coefficient is significantly statistically different from zero at the 0.10, 0.05 and 0.01 levels, respectively. SES denotes socioeconomic status.