Economics Bulletin

Volume 38, Issue 2

Decomposing the language pay gap among the indigenous ethnic minorities of Mexico: is it all down to observables?

Adriana Aguilar-Rodriguez
Center for Research in Geospacial Information Science (CentroGeo) & PANEL, Mexico

Alfonso Miranda
Economics Division & PANEL, Centre for Economic Research and Teaching (CIDE), Mexico

Yu Zhu
Economic Studies, University of Dundee

Abstract

Using the decomposition methods of Oaxaca and Choe (2016), we investigate the pay gap between indigenous language monolinguals (INL) and Spanish-indigenous-language bilinguals (BIL) among indigenous ethnic minorities in Mexico using the 10% sample of the Mexican Census 2000 and 2010. The decomposition fits linear models with municipal fixed effects for the case of males and correlated random effects Heckman sample selection models for the case of females (to account for potential sample selection bias). We find evidence of a positive return to bilingualism for males of 17% and of 42% for females. Over 60% of the pay gap is explained by differences in observable characteristics.

PANEL is the Program for Longitudinal Studies, Experiments and Surveys of CIDE. Acknowledgement: We would like to thank Ronald Oaxaca and Chung Choe for providing us with a copy of their Stata code for implementing their panel decomposition method. We are also grateful to Quetzali Ramirez-Guillen, Valeria Gracia-Olvera, Cristina Alvarez-Venzor and Javier Valbuena for their excellent research assistance. Funding: This work was supported by CONACYT repatriation program for Alfonso Miranda [grant number 191523]. Conflict of Interest: We declare not to have any conflict of interest.
Citation: Adriana Aguilar-Rodriguez and Alfonso Miranda and Yu Zhu, (2018) "Decomposing the language pay gap among the indigenous ethnic minorities of Mexico: is it all down to observables?", *Economics Bulletin*, Volume 38, Issue 2, pages 689-695
Contact: Adriana Aguilar-Rodriguez - aaguilar@centrogeo.edu.mx, Alfonso Miranda - alfonso.miranda@cide.edu, Yu Zhu - yuzhu@dundee.ac.uk.
Submitted: February 06, 2018. **Published:** April 15, 2018.

1. Introduction

Learning a second language is a costly human capital investment motivated by the possibility of economic returns (Chiswick and Miller 2003; 2014). Most of the existing literature on economics of language looks at deficiency in host country language for international migrants (Chiswick and Miller 2014). Only a couple of papers have looked at the returns to proficiency in the dominant language among ethnic minorities, see Chiswick et al. (2000), Godoy et al. (2007). The present paper intends to contribute to this literature by offering evidence on how large the language pay gap is among the ethnic minorities of Mexico; how much of that gap is stripped away once the effect of demographics and time-invariant municipal effects are accounted for by regression analysis; and finally, what proportion of the raw gap is explained by differences in observable characteristics across language groups and what proportion of the gap remains unexplained — potentially due to labor market discrimination. From the point of view of the authors, this initial study may help open the literature on the effects of indigenous language in the Mexican labor market: an important topic that remains untouched.

We use data from the 10% sample of the Mexican Census 2000 and 2010, which contain information about indigenous language and self-attribution to an ethnic minority. According to the Census there were 97 million Mexicans in 2000, of which 5.3 million (5.5%) considered themselves as belonging to an indigenous minority by culture. In 2000 nearly 90.7 million Mexicans (93.5%) reported to speak Spanish exclusively, whereas 1.2 million (1.2%) reported to speak only an indigenous language and 5.1 million (5.3%) to be bilingual. The language demographics remained almost unchanged 10 years after.

To ensure that we have a suitable comparison group, the analytic sample only uses data from individuals who self-identify as indigenous by culture and who are either indigenous language monolinguals (INL) or Spanish-indigenous-language bilinguals (BIL).

The analytic sample contains only prime-aged individuals (aged between 20 and 40) so that we look at differences in pay at the peak of the working life. Students, permanently disabled (for work), and retirees are excluded. The sample contains $888,885$ individuals, 53.7% females and 46.3% males. Table 1 presents descriptive statistics. The outcome variable is (monthly) log-income, which is measured in 2010 constant pesos. The main independent variable is an indicator for bilingual BIL speaker (control group is indigenous language monolingual INL). At individual level other controls include age, age squared, education, Catholic religion, social class, dwelling's number of rooms, and materials in walls and ceilings as controls for wealth. At the settlement level (village) we control for population size. At the municipal level we control for population, population squared, municipal international migration rate, and inter-municipal domestic migration rate. Finally, we add a year dummy.

2. Estimation strategy and sample selection

We use linear models to obtain a regression-adjusted language pay gap, by gender, using individual level data. The dependent variable is $logw_{imt} = log(\text{income}_{imt})$, with $i = \{1,\ldots,N\}$ individuals,

Table 1. Descriptive statistics of the analytic sample

Var	Mean	Description
lincome	7.45	log of income
work	0.52	Work status
BIL	0.82	bilingual
female	0.54	female
age	29.49	age
cathoik	0.76	Catholic
noedu	0.18	no education
primary	0.56	primary
secry	0.18	secondary
prep	0.08	preparatory
tamloc1	0.75	locality pop. $\leq 2.5k$
tamloc2	0.2	$2.5k >$ locality pop. $\leq 15k$
tamloc3	0.02	$15k >$ locality pop. $\leq 100k$
tamloc4	0.02	locality pop. $> 100k$
intmig	0.00	mun. international mig. rate
munmig	0.15	mun. domestic mig. rate
pop	4.13	mun. population 10k
walloth	0.11	walls other
wallclay	0.19	walls clay
floctre	0.61	floor concrete
flowood	0.04	floor wood
ceiloth	0.21	ceil other
ceilmetal	0.46	ceil metal
ceiltile	0.09	ceil tile
nroom	2.64	# rooms
ssclfarm	0.07	SES farm
ssclsfempl	0.31	SES employee
nprimch	0.71	# pre-school & primary school children

Note. $N = 888,885$ except for log income where $N = 265,904$.

$m = \{1, \ldots, M\}$ municipalities, and $t = \{2000, 2010\}$ years. We start by pooling the two cross-sections and fitting

$$logw_{imt} = \mathbf{x}_{imt}\beta + \theta BIL_{imt} + \mathbf{w}_{mt}\gamma + \delta d_{2010} + u_{imt}, \tag{1}$$

by pooled OLS (POLS) regression. $\mathbf{x}_{imt}$ represents individual controls other than language, $\mathbf{w}_{mt}$ represents municipal controls, BIL_{imt} is a dummy for bilingualism, and d_{2010} is a year dummy. The parameter of interest is θ. As starting point, we drop missing income observations and assume that data are missing at random (MAR) (see Little and Rubin 2002). Clustered standard errors at municipality level are used for inference.

Economic theory suggests that much of the variation of income among workers who live in different places is explained by systematic differences in the local labour market conditions that they experience, which ultimately is unobserved heterogeneity. To account for this, we introduce a municipality fixed-effect c_m and fit

$$logw_{imt} = \mathbf{x}_{imt}\beta + \theta BIL_{imt} + \mathbf{w}_{mt}\gamma + \delta d_{2010} + c_m + u_{imt}, \tag{2}$$

by linear municipal fixed-effects. An alternative is to use a correlated random effects (CRE) estimator that models c_m as a function of $\mathbf{w}_{mt}$ as well as (municipality) mean individual controls $\bar{\mathbf{x}}_m = (N_m T)^{-1} \sum_{i \in m} \sum_t \mathbf{x}_{imt}$ fitting

$$logw_{imt} = \mathbf{x}_{imt}\beta + \theta BIL_{imt} + \bar{\mathbf{x}}_m\gamma_1 + \mathbf{w}_{mt}\gamma_2 + (\mathbf{w}_{mt} \times d_{2010})\,\gamma_3 + \varepsilon_{imt}, \tag{3}$$

by POLS. Clustered standard errors at municipality level are used for inference.

Over 87% of the indigenous men aged $20-40$ in Mexico work, which is practically the same rate as the mestizo men. At such high levels of labor market attachment no major issues of sample selection are anticipated in our analysis for men. In contrast, only 34% of the indigenous females aged 20 to 40 work. We address potential sample selection bias in our female regressions by implementing Wooldridge (1995)'s correlated random effects (CRE) sample selection estimator with fixed-effects at the municipality level (see online appendix). As instrument for selection we use the number of pre-school and primary school children in the household, which we postulate to affect a woman's probability to work but not directly her wage once at work. This variable is often used in the literature as instrument for female labour market participation (see Heckman and Macurdy 1980, Mroz 1987).

3. Main results

The first two columns of table 2 compare bilingual males to indigenous language monolingual males. This specification gives descriptive evidence of what are, for an indigenous language monolingual, the returns of learning Spanish. POLS results indicate that going from INL to BIL increases log-wage by 0.22 log-units, net of ethnicity, age, education, religion, and controls at municipal level that include population, and rate of international and domestic migration. This effect is statistically significant at the 1% and equivalent to a positive return to bilingualism of about 24%.

Table 2. Regressions for log(income)

	Males		Females		
	POLS[a]	FE[a]	POLS[a]	FE[a]	CRE Heckman[b]
	(1)	(2)	(3)	(4)	(5)
bilingual	0.218***	0.160***	0.423***	0.280***	0.353***
	(0.026)	(0.012)	(0.034)	(0.024)	(0.028)
invMills					-0.218***
					(0.052)
invMills×2010					0.014
					(0.025)
No. primary school children 2000					-0.025***
					(0.006)
No. primary school children 2010					-0.047***
					(0.005)
F for exclusion of instrument 2000					12.373***
					(3.228)
F for exclusion of instrument 2010					75.307***
					(9.858)
N. of obs	193,241	193,241	72,571	72,571	477,127
N. of clusters	1,953	1,953	1,648	1,648	2,033
R^2-Adjusted	0.28	0.37	0.38	0.49	0.42

Note. *10% significant; **5% significant; ***1% significant. Individual controls: age, age^2, religion, education, size of locality, dwelling's building materials, number of rooms, social class. Locality controls: locality size. Municipality controls: population (continuous), population squared, rate of international migration, rate of intermunicipal migration. The same controls are used for males and females. For female's selection, we use number of pre-school and primary school children as instrument.

[a] Municipality clustered robust standard errors in parenthesis.

[b] Estimates from Wooldridge's correlated random effect's (CRE) Heckman sample selection estimator. Bootstrap municipality clustered standard errors in parenthesis (50 repetitions).

The finding is qualitatively consistent with the results reported by Godoy et al. (2007) for Bolivia: bilingualism has a positive return when compared to indigenous language monolingualism. Fitting the model by municipality fixed effects in column 2 reduces the size of the coefficient to 0.16, which is equivalent to a positive return to BIL of 17%.

Column 3 of table 2 shows that females get a positive 0.42 log-unit return to bilingualism after controlling for ethnicity, age, education, religion, and controls at municipal level. This is equivalent to a 52% income return that is statistically significant at 1%. Fitting the model by fixed effects at the municipal level brings down the effect to 0.28 log-units. Finally, accounting for potential sample selection bias and fitting the CRE Wooldridge (1995)'s estimator shows that bilingualism carries a 0.35 log-units coefficient that is statistically significant at 1% and equivalent to a 42% positive return to bilingualism for females.

4. Oaxaca-Choe decomposition

Oaxaca and Choe (2016) implement a panel decomposition of the log-wage gap on the basis of Wooldridge (1995)'s CRE sample selection estimator, which is a longitudinal data extension of the Heckman (1979) two-step estimator. As in standard Oaxaca (1973) and Blinder (1973), the aim is to uncover what proportion of the log-wage gap between two groups is explained by differences in observable characteristics (the 'E' part) and what proportion is left 'unexplained' once the effect of observables is netted out (the 'U' part). Oaxaca and Choe extend the usual toolkit to (a) allow the two groups to differ in their labor market attachment and (b) to control for unobserved heterogeneity at the panel level.

Depending on whether the effect of selection and/or the effect of unobserved heterogeneity are considered as 'explained' or 'unexplained', Oaxaca and Choe define six decomposition methods. In the context of a repeated cross-section, we choose to focus on Method 1 which assumes that the 'explained part' is anything due to differences in characteristics between groups, and that the 'unexplained part' is anything that is due to differences in parameters. Differences in c_m as well as differences in inverse Mills ratio terms are therefore part of the explained part.

Table 3 presents results. Bootstrapped municipality clustered standard errors (50 repetitions) are reported in parentheses. In all cases, indigenous language monolinguals are the 'reference' or 'control' group. For males, we find that 61% of the pay gap is explained by differences in observed characteristics, while 39% of the language gap is unexplained — potentially due to labor market discrimination. As for females, we find that 62% of the gap is due to differences in observed characteristics and 38% is due to differences in parameters.

5. Conclusions

Using the decomposition methods of Oaxaca and Choe (2016), we investigate the pay gap between indigenous language monolinguals (INL) and Spanish-indigenous-language bilinguals (BIL). We aim to offer evidence on how large the language pay gap is among the ethnic minorities of Mexico; how much of that gap is stripped away once the effect of demographics and time-invariant munic-

Table 3. CRE Oaxaca & Choe log(income) decomposition

	Males		Females	%
$\overbrace{log(\text{income})_{\text{BIL}} - log(\text{income})_{\text{INL}}}^{\text{raw pay gap}}$	0.446***		0.936***	
	(0.042)		(0.076)	
Explained	0.273***	61%	0.581***	62%
	(0.032)		(0.058)	
Unexplained	0.173***	39%	0.355***	38%
	(0.019)		(0.028)	
# obs	397,797		469,273	
# of clusters	2,006		2,026	

Note. *10% significant; **5% significant; ***1% significant. Bootstrap municipality clustered standard errors in parenthesis (50 repetitions). Controls included: see Table 2 note.

ipal effects are accounted for by regression analysis; and finally, what proportion of the raw gap is explained by differences in observable characteristics across language groups and what proportion of the gap remains unexplained. Individual level pooled cross-section data from the 10% sample of the 2000 and 2010 Mexican Census are used. For males, linear models with municipal fixed effects are fitted to estimate the language pay gap net of ethnicity, age, education, religion, population and migration at municipal level. To account for potential sample selection bias, Wooldridge (1995)'s CRE Heckman sample selection estimator is fitted for females. Oaxaca and Choe (2016)'s pay gap decompositions are performed.

Findings show that there is a substantive positive return to bilingualism in Mexico (17% for males and 42% for females). Most of the language pay gap is explained by differences in observable characteristics (around 61% for males and 62% for females).

References

Blinder, A. S. (1973) "Wage discrimination: Reduced form and structural estimates" *The Journal of Human Resources* **8**,436–455.

Chiswick, B. and Miller, P. W. (2003) "The complementarity of language and other human capital: Immigrant earnings in canada" *Economics of Education Review* **5**, 469–480.

Chiswick, B. and Miller, P. W. (2014) "International migration and the economics of language" In Chiswick and Miller, editors, *Handbook on the Economics of International Migration*. Elsevier.

Chiswick, B., Patrinos, H., and Hurst, M. (2000) "Indigenous language skills and the labor market in a developing economy: Bolivia" *Economic Development and Cultural Change* **2**, 349–67.

Godoy, R., Reyes-Garcia, V., Seyfried, C., Huanca, T., Leonard, W., McDade, T., Tanner, S., and Vadez, V. (2007) "Language skills and earnings: Evidence from a pre-industrial economy in the bolivian amazon" *Economics of Education Review* **3**,349–360.

Heckman, J. J. (1979) "Sample Selection Bias as a Specification Error" *Econometrica*, **47**,153–161.

Heckman, J. J. and Macurdy, T. E. (1980) "A Life Cycle Model of Female Labour Supply" *The Review of Economic Studies* **47**,47–74.

Little, R. J. A. and Rubin, D. B. (2002) *"Statistical anaysis with missing data"* Wiley, Hoboken, NJ.

Mroz, T. A. (1987) "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions" *Econometrica* **55**, 765–799.

Oaxaca, R. (1973) "Male-female wage differentials in urban labor markets" *International Economic Review* **14**,693–709.

Oaxaca, R. and Choe, C. (2016) "Wage decompositions using panel data sample selection correction" *Korean Economic Review* **32**, 201–218.

Wooldridge, J. M. (1995) "Selection corrections for panel data models under conditional mean independence assumptions" *Journal of Econometrics* **68**, 115–132.