I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17741

The Causal Effect of Speaking Spanish as an Additional Language on Education, Labor, and Wellbeing Outcomes, Among the Indigenous Ethno-Linguistic Minorities of Mexico

Alfonso Miranda

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17741

The Causal Effect of Speaking Spanish as an Additional Language on Education, Labor, and Wellbeing Outcomes, Among the Indigenous Ethno-Linguistic Minorities of Mexico

Alfonso Miranda
CIDE and IZA

FEBRUARY 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Causal Effect of Speaking Spanish as an Additional Language on Education, Labor, and Wellbeing Outcomes, Among the Indigenous Ethno-Linguistic Minorities of Mexico*

We estimate the effect of speaking Spanish as an additional language (SAL)—as opposed to speaking it as a native—on education, labor, and wellbeing outcomes among Mexico's indigenous ethno-linguistic minorities. Controls are appropriately comparable indigenous individuals who speak only Spanish. To address treatment endogeneity, we use 2SLS, maximum likelihood, and control function estimators, using parental indigenous language status as instruments. Unlike prior studies, we account for key confounders: parents' education, occupation, and imputed family income. SAL reduces education by one year—equivalent to a–0.2 standard deviations reduction on schooling. It also imposes a–23% penalty on permanent income (wealth). No significant effects are found on employment, life satisfaction, or health.

JEL Classification: J71, J31, 12, 128

Keywords: Spanish as additional language, indigenous ethno-linguistic minorities, economic outcomes, Mexico

Corresponding author:
Alfonso Miranda
Center for Research and Teaching in Economics (CIDE)
Region Centro campus
Circuito Tecnopolo 117
Aguascalientes 20313
Mexico.
E-mail: alfonso.miranda@cide.edu

* Competing interests statement: None. Role of the funding source: None Declarations of interest: None. Hazards and Human or Animal Subjects: The present study uses secondary data. Acknowledgement: I am grateful to PROCER research team who originally designed and collected the survey data analyzed in the present study. I am also grateful to Ana Isabel Corona Pantoja for her excellent work as research assistant. I dedicate this work to my late mother and father, whose unwavering support and encouragement made my academic journey possible.

1. Introduction

In the present paper we aim to study what is the effect of speaking *Spanish as an additional language* (SAL)—as opposed to speaking it as a native—on education, labor, and wellbeing outcomes, among the indigenous ethno-linguistic minorities of Mexico. Controls are appropriately comparable indigenous individuals who speak only Spanish. We expect to find an approximation of the penalty for not speaking the language natively—i.e. with some deficiencies and/or accent. Beyond exploring the sign and magnitude of partial correlations, we intend to estimate true causal effects.

The literature has fundamentally studied how migrants who do not speak the language of the host/destination country become bilinguals and adapt to the local labor market in their process of cultural and economic assimilation (see, for instance Chiswick 2009, Chiswick and Miller 2018). Less is known about how native ethno-linguistic minorities that speak the language of the majority as an *additional language* (AL) fare in the labor market in their process of cultural and economic assimilation to their 'reference majority.'

Our task is greatly complicated by the fact that language cannot be credibly experimented upon: It takes years to learn and comes intimately linked with ethnicity, culture, geography, and socio-economic status. As soon as a person speaks a complex mixture of cues, impossible to fake, are given to the listener: accent, syntax, semantics. Such cues allow easy speaker social identification that frustrates any attempt to vary language experimentally. Also, though there are some examples, pseudo-experimental interventions that affect individuals' language without directly affecting education, labor, and/or wellbeing outcomes, are difficult to come by. The introduction of Catalan language requirements for performing public jobs in Catalonia is one, rare, example. For most countries and languages, however, such sources of pseudo-experimental variation are unavailable. The only hope to identify a language causal effect must, then, rely on the use of instrumental variables. We take such avenue, knowing the difficulties and challenges we face to produce credible estimates.

To ameliorate concerns coming through the action of possible confounders, we use data from the Ethno-Racial Discrimination in Mexico Project Survey 2019 *(Proyecto sobre Discriminación Étnico-Racial en México)* (PRODER 2019), which we argue are particularly suited to study the topic. PRODER is unique because sufficiently oversamples Mexican indigenous ethno-linguistic minority groups and collects a set of rich individual level characteristics, including self-reported ethnicity and linguistic family heritage. This gives us the opportunity of using parents' language skills as instruments for individuals' own language skills on regressions of our outcome variables on a (binary) *Spanish as an additional language* (SAL) indicator. To secure a control group as clean as possible we only use data from individuals who self-declare to be indigenous—as non-

indigenous individuals are in general too different to be a suitable control group. Instruments are strong predictors of treatment status and, as we discuss later in detail, over identification tests suggest they are uncorrelated with our responses' error terms—i.e. we have reasons to believe instruments are valid. Moreover, unlike previous work, we are able to control for potential key confounders that are rarely observed, including mother's education, father's education, mother's occupation, father's occupation, and origin family imputed permanent income. Also, we control for an explicit measure of individuals' skin tone; which in the ethno-racial discrimination literature has been suggested to better approximate "ethnicity" in Latin American countries, where race and ethnicity are ambiguous due to the *meztizaje* (mixing) process that have existed since colonial times (see, for instance, Solís et al. 2025, Telles et al. 2015, Telles and Paschel 2014, Flores and Telles 2012). This gives us confidence that we are estimating true causal effects.

Findings suggest that indigenous individuals who speak Spanish as an additional language in Mexico complete one year of education less than the suitable comparable group of Spanish monolinguals. This effect is equivalent to a -0.2 standard deviations drop on schooling. Also, we find SAL carries a penalty of about -23% on a measure of permanent income (or wealth); equivalent to an income cut of about -0.51 standard deviations. On another measure of wealth, number of bulbs in the household, the estimated treatment effect is -2.1; equivalent to a drop of about 30%. No statistically significant SAL effects are found on current work status, life satisfaction, or health status.

Our contributions are threefold. First, we contribute a piece of applied work in an emerging and under explored subfield concerned with studying how a lack of native command of the majority's language may cause important economic and welfare penalties among native ethno-linguistic minorities (De la Fuente Stevens and Pelkonen 2023, Godoy et al. 2007a, Chiswick et al. 2000a, Patrinos et al. 1994, are the few other pieces in the subfield). Second, unlike most previous work in the subfield, we explicitly address the problem of potential treatment endogeneity. In fact, in our attempt to deal with treatment endogeneity, we contribute to the general economics of language literature by exploring the use of family linguistic heritage—a rarely observed variable—as instrument for an individuals' own language skills in different wellbeing response variables. We find evidence suggesting that the strategy works well when sufficient family background controls are available. Finally, we contribute a case study that is important and under researched. Mexico has a large and diverse population of indigenous etno-linguistic minorities that speak Spanish as an additional language (SAL)—over 7 million who speak 68 different languages according to the 2020 Census. Despite the relevance of the case study, and long-standing anecdotal accounts that SAL may cause economic and social disadvantage among Mexican minorities, too little work on

the topic has been done in the past. De la Fuente Stevens and Pelkonen (2023), Aguilar-Rodriguez et al. (2018) and Parker et al. (2005) are the only previous pieces of related work. We improve upon these studies by explicitly focusing on the SAL effect, using better quality data for analysis, and addressing the potential problem of treatment endogeneity. While Mexico is an interesting case in its own right, it shares a common history and a similar socio-economic structure with other Latin American countries. Studying the Mexican case is, therefore, likely to provide insights that may be relevant elsewhere.

2. Related literature

The economics of language studies what returns an individual who belongs to an ethno-linguistic minority can gain by learning to speak the language of the majority (Chiswick 2009, Chiswick and Miller 2018). Speaking the language of the majority becomes a form of human capital because: (i) it is costly to acquire (requires investment in time, effort, and money), (ii) once learned, becomes a skill inherent to the individual, and (iii) it is productive because it eases communication at the marketplace, at school, at the health center, at the judiciary, and other important state and civil institutions; hence, it is expected to improve labor market outcomes (v.g. probability of employment and/or income) as well as welfare. Besides pursuing pecuniary returns, individuals form ethno-linguistic minorities may choose to learn the majority's language to gain purely non-pecuniary returns (Grenier et al. 2021). Speaking the language of the majority, for instance, can help an individual to develop a sense of belonging, become better accepted by the community, and to establish healthy and nurturing personal relationships—which do not derive in pecuniary gains—with people beyond her/his own minority ethno-linguistic group.

Learning the language of the majority is a continuum that goes from monolingualism on the minority's language to speaking the majority's language as a native. A bilingual is an individual who reaches a point that allows her to successfully communicate in the language of the majority to perform everyday activities that require use of the skill (Weinreich 2010, Grosjean 2010).[2] Perfect bilingualism, i.e. speaking the two languages as a native, is rarely achieved (Grosjean 2010, p. 20). In most cases, even if communication is never broken, there are signs, or cues, that allow

[2]Despite great effort over nearly 100 years of research, linguists have find it difficult to define 'bilingualism'. Early definitions put emphasis on proficiency and defined bilingualism as "the native control of two languages" (Bloomfield 1933, p. 56). More recently the emphasis shifted into usage, defining bilingualism as "The practice of alternatively using two languages will be called [here] bilingualism, and the person involved bilingual" (Weinreich 2010, p. 1). Here we refer to bilingualism in his modern meaning, where 'bilingual' refers to the regular use of two languages, not that the two languages are spoken perfectly or with no accent.

identification: accents, grammar construction, vocabulary (see, for instance Flege 1987, Johnson and Newport 1989, Meara 1980, among others). Such cues set apart native speakers from individuals who speak the majority's language as an additional language; a sort of intermediate point. In this continuum economic theory hypothesizes that moving from monolingualism on the minority's language to speaking the language of the majority—first as an additional language, then as a native—carries important labor and welfare returns (see, for instance Chiswick and Miller 1995; 2015).

The economics of language is a diverse field. Most of the available studies focus on investigating the relationship between language and employment probability and wage. More recently, however, there is increasing interest on investigating its potential effects on education, health, and generally, welfare. Different studies set out to estimate diverse effects depending on the study population and the institutional context. Treatment and control groups across sub-strands of the literature are often different and non-comparable. As a consequence, the expected sign of the effect is sometimes positive and sometimes negative. This often creates confusion as findings look mixed at first sight, when in fact most of the available evidence fit well together and form a coherent body of knowledge.

The main strand of the literature use host/destination country language proficiency in the context of international migration. Proficiency is often categorical and self-reported. A measure that suffers from substantial measurement error (Dustmann and Soest 2001). Theoretically the treatment is continuous as proficiency varies at individual level. The control group are members of the immigrant population who are or remain minority (foreign) language monolinguals after arrival to the destination country. We call this the "majority language proficiency" (MyPL) treatment. Unobserved heterogeneity is often a concern, as individual cognitive ability is likely to affect both language skills and labor market outcomes—a variable that in most cases is unobserved by the researcher. This induces a potential problem of treatment endogeneity (Chiswick and Miller 1995), which is addressed by IV, matching or DiD. Popular instruments include interview language (Dustmann and Fabbri 2003), an interaction between age at arrival and country of birth (Bleakley and Chin 2004), lingustic distance (Donado 2017), and language proficiency of family members (Wang et al. 2017). The expected sign of the treatment is positive as better proficiency in the host country language should improve immigrants' labor market outcomes and welfare. Chiswick and Miller (2015) provide an excellent review of the literature.

Correcting for potential treatment endogeneity Chiswick and Miller (1995) report that immigrants who are fluent in the language of the majority in the host country earn higher wages: 16% for the USA, 12% for Canada, 9% for Australia, and 11% for Israel. For the UK, correcting for

endogeneity and measurement error, Dustmann and Fabbri (2003) find that fluency in English increases employment probability by 22% and wages by 18% – 20%. For Germany, Dustmann and Soest (2001) find that a standard deviation increment in their measure of German fluency raises wages by about 7%. Importantly, evidence shows that individuals overreport rather than underreport language proficiency, and that the downward OLS mesurement error bias dominates the OLS upward endogeneity bias. So, after correcting for both sources of bias the effect of language on wages becomes more positive than in OLS estimates. In the USA, using an interaction between age at arrival and country of birth as instrument to correct for treatment endogeneity, Bleakley and Chin (2004) find that migrants who speak English well and very well earn 33% and 67% higher wages than those who do not speak English at all. In Australia, also correcting for treatment endogeneity, Guven and Islam (2015) find that moving from low to high English proficiency increases wages by about 35%. In the Netherlands, Yao and Van Ours (2015) fit an IV estimator using age of arrival and country of origin as instrument to correct for treatment engogeneity and measurement error, and find that immigrants with Dutch proficiency earn wages 48% higher to immigrants with Dutch language problems. For Israel, Berman et al. (2003) studies the case of Soviet immigrants using longitudinal data with a time-varying measure of Hebrew proficiency. Fitting a first-difference estimator the authors find a 5.7% increase in wages for each unit of Hebrew proficiency on a four step scale. Other important studies include McManus (1985), Dustmann and Van Soest (2002), and Shields and Price (2002).

A second strand of the literature considers the case of native ethno-linguistic minorities who speak a minority language/dialect and move to speak the language of the majority as a second language. The treatment is in most cases binary. The control group are individuals who remain minority language/dialect monolinguals. We call this the "Bilingualism" (BIL) treatment. For Canada, fitting OLS regressions with a rich sets of controls, Grenier (1987) finds a 5% wage premium for francophones who speak English as an additional language, and a 10% premium for anglophones who speak French. In China, after correcting for treatment endogeneity, Gao and Smyth (2011) reports a 42% wage premium for domestic migrants in China who speak Mandarin (the language of the majority), relative to migrants that speak a minority dialect/language. In the Netherlands, fitting pooled OLS regressions, Yao and van Ours (2019) find a 8% (females) and 10% (males) wage premium for speaking standard Dutch rather than a Dutch dialect. In the USA, Grogger (2019) looks at speech patterns among American English native speakers, distinguishing individuals who speak with a distinctive African American Vernacular English (AAVE) or a distinctive Southern American English (SoAE) as opposed to speaking with Standard American English (SAE) accent. No attempt for correcting treatment endogeneity and/or measurement error is done. He finds that

Black individuals with mainstream speech (SAE) earn 13.6% more than blacks with AAVE accent. And southerners with mainstream (SAE) accent earn 8.6% more than southerners with SoAE accent. In Mexico, among individuals who declare to be indigenous, Aguilar-Rodriguez et al. (2018) find that bilinguals who speak an indigenous language and Spanish earn 52% (males) and 41% (females) more than individuals who are indigenous language monolinguals—no correction for treatment endogeneity and/or measurement bias is attempted. In China, focusing in health status as response variable, Lu et al. (2019) study the case of different cohorts of domestic migrants that move within and across different dialect regions. Using a difference-in-difference estimator, the authors find that elderly migrants who do not cross a dialect border report a health status that is 0.024 standard deviations higher than elderly migrants who do cross a dialect border.

A third strand of the literature, the closest related to the present work, is concerned with target populations that already speak the majority's language as an additional language, as it is the case of long-term migrants, second generation migrants, and native ethno-linguistic minorities. Here the focus changes to learn what are the potential economic and welfare penalties that these populations face for not achieving full majority's language native proficiency to assess whether diversity and anti-discrimination policies are needed and justified. In such a context, speaking the majority's language as an additional language (MyAL)—as opposed to speaking it as a native—becomes the treatment of interest. The control group are individuals from the ethnic minority who speak the language of the majority as natives, often (though not necessarily) after loosing the ability to speak their minority language. In this case the expected sign of the effect is negative as speaking with deficiencies the language of the majority is anticipated to harm labor market prospects. We call this the "majority language as additional language" (MyAL) treatment effect. Notice that, because treatment and control group are different, BIL and MyAL are related but different.

Grenier (1984) studies the Hispanic-American native population fitting OLS regressions. No correction for treatment endogeneity bias or measurement error is implemented. The author finds that speaking English with deficiencies reduces wages by -14%. Also for the Hispanic-American population, Gonzalez (2005) obtains non-parametric bounds that correct for endogenous treatment bias using age of arrival as monotone instrumental variable and finds than limited English proficiency (LEP) leads to a wage penalty that lies between -3.8% and -38.6%. Also, LEP is found to reduce the probability of employment in a range that goes from 0 to -6.5 percentage points. Another important study of the native ethno-linguistic minorities that speak English as an additional language in the USA is Chiswick and Miller (2016), which uses data from the $2005-2009$ American Community Survey. OLS regressions by language groups are fit. No correction for treatment endogeneity or measurement error is attempted. According to the study, bilinguals earn -4.7%

lower wages than monolingual English speakers, but there is strong heterogeneity by language spoken. Spanish speakers earn -20% less than the monolingual English speakers, while speakers of some Western European, East Asian, and Hebrew, languages earn significantly higher wages than monolingual English speakers. For the UK, Miranda and Zhu (2013) use an IV estimator with age at migration interacted by origin country as instrument. The authors find that English deficiency has a causal negative effect of -23% on wages. In a developing country context, Chiswick et al. (2000b) studies the effect of speaking an indigenous language in Bolivia, comparing Spanish monolinguals and Spanish-Indigenous language bilinguals. Fitting OLS regressions that do not control for potential endogeneity and/or measurement error bias, the authors find bilinguals who speak Spanish as a second language earn wages that are -23% (males) and -28% (females) lower than Spanish monolinguals. While for men language does not affect labor market participation, bilingual women are 7 percentage points more likely to work than Spanish monolinguals. Looking at unemployment duration among the Russian-speaking minorities in three Estonian regions with varying ethnic concentrations, Lindemann (2014) fits hazard models that compare Russian speakers with poor Estonian proficiency and Russian speakers with good Estonian proficiency. The author finds that those with Estonian deficiency move out of unemployment with lower probability.

Sometimes authors take minority language monolinguals as control group and put together majority language monolinguals and bilinguals who speak the majority language in the treatment group. This defines a sort of "majority language" treatment (MyL) effect. This is, however, not very popular because majority language monolinguals are often very different to minority language monolinguals in many observable and unobservable characteristics other than language; and hence not a good comparison group. Examples of this include De la Fuente Stevens and Pelkonen (2023), who use a matching estimator and find that speaking Spanish in Mexico is associated with a 29% increase on earnings in Mexico among the population who self-declare to be indigenous by culture. Also for Mexico, comparing Indigenous language monolinguals (treatment) to Spanish monolinguals plus bilinguals (control), Parker et al. (2005) fit OLS regressions with a rich sets of controls, and find that a mother who is indigenous language monolingual is associated with a reduction on a child's school enrollment of -2.2 percentage points and an schooling attainment reduction of 0.24 years among those aged between 6 and 18. Another example is Godoy et al. (2007b) who studies the case of Bolivia and finds that Spanish speakers, which include Spanish monolinguals and bilinguals, earn between 36.9 and 46.9% more than monolingual speakers of the local language. Patrinos et al. (1994) analyzes data from Asuncion, Paraguay, where 56% is Spanish-Guarani bilingual while 36% are Spanish bilinguals—only 8% are Guarani monolinguals.

Fitting OLS regressions, they find that, controlling for education and years of experience, Guarani speakers earn wages that are -11% lower than Spanish speakers. Nearly 79% of the wage differential is due to differences in observables, while 21% is due to differences in observables. In these studies no correction for treatment endogeneity or measurement error bias is attempted.

3. Data

According to the Mexican Census 2010 there were 110.6 million Mexicans. A total of 15 million (14%) declared to be indigenous by culture, while 95.1% million (85%) declared not to be indigenous. Out of the 15 million indigenous, 9 million (58%) are Spanish language monolinguals, 1.2 million (8%) are indigenous language monolinguals, and 5.2 million (34%) are bilinguals.[3] Ten years after, the Census 2020 found similar ethno-linguistic demographics.

These demographics imply that, unless a sample is explicitly designed to study indigenous minority groups, general propose surveys are unfit to study the effect of language on wellbeing in Mexico—as it neither draw representative samples nor sufficient observations of Mexico's indigenous minorities. Oversampling is, therefore, required.

To overcome the challenge, the present paper analyses individual level data from the Ethno-Racial Discrimination in Mexico Project Survey 2019 *(Proyecto sobre Discriminación Étnico-Racial en México)* (PRODER 2019). PRODER is an innovative cross-section survey designed by a multidisciplinary research team at Colegio de México (Colmex) with the aim of studying how racism and ethnoracial perceptions of discrimination relate to socioeconomic inequality in Mexico. Key to our propose, as well as drawing a national representative sample, PRODER explicitly oversamples areas with high indigenous populations in Oaxaca, Merida, and the Maya area.[4] The target population are 25 to 64 Mexican individuals living in households between July 30th and October 11th 2019. Computer assisted personal interviewing (CAPI) was implemented—mainly using smart phones. Interviews were all done in Spanish (further details on the sample and questionary design are offered in the online appendix).

Besides oversampling indigenous minority groups, PRODER is unique because it collects information about linguistic family heritage along with self-reported ethnicity. This makes it particularly suitable to explore the relationship between an individual's ability to speak Spanish as additional language and her/his wellbeing, as parents' language skills can be used as instrumen-

[3]There are a total of 68 Indigenous languages in Mexico. Among the most spoken are: Náhuatl, Maya, Tseltal, Tsotsil, and Zapoteco. Together these 5 languages represent 47% of the population who speak an Indigenous language.

[4]This region includes the muncipalities of: Hopelchén, Calakmul, José María Morelos, Cantamayec, Chacsinkín, Chankom, Chikindzonot, Maní, Mayapán, Ozkutzcab, Tahdziú, Teabo, Tekom, Tixcacalcupul, Tixmehuac y Yaxcabá.

tal variables. As a consequence, the researcher can go beyond estimating partial correlations and venture to try estimating causal effects.

To secure a control group as clean as possible we only use PRODER data from individuals who self-declared to be indigenous (*ineth*=1), as non-indigenous individuals are unsuitable as control group—i.e. too different in many observable and unobservable characteristics. 32% of PRODER's $N = 7,187$ sample comply with such condition. To secure adequate overlap between the sample who speaks an indigenous language $(inlang = 1)$ and the sample who does not $(inlang = 0)$, we further eliminate data from states where there are too few individuals who declare an indigenous language. These leave us with data from 7 Mexican states where most of the indigenous population concentrates: Yucatán, Chiapas, México, Michoacán, Oaxaca, Quintana Roo and Veracruz. After applying these exclusions the analytical sample has $1,300$ observations: 686 females (52.7%) and 614 males. Table 1 offers descriptive statistics.

3.1. Response variables

All PRODER interviews were done in Spanish. Hence, all individuals in the sample who speak an indigenous language speak Spanish as well (i.e. they are bilinguals). As a consequence, we define individuals who speak *Spanish as additional language* (SAL) as the set of individuals who declare speaking an indigenous language in the PRODER sample. We study the relationship between SAL and a series of individual wellbeing indicators: (i) current labor market participation (*work*), (ii) years of completed education (*yrsedu*), (iii) imputed permanent income or wealth (*iincome*), (iv) number of bulbs in the household (*bulbs*) as a proxy for income, (v) a measure of life satisfaction (*lifesat*), and (vi) a measure of health status (*hthstat*).

work is simply a dichotomous variable that takes on 1 if the individual currently works, with mean 0.69 ($SD = 0.46$). *yrsedu* is a non-negative integer variable with mean 11 ($SD = 5$) and five modes that reflect the different levels of education in Mexico.[5] To simplify the analysis, we treat *yrsedu* as a continuous variable. A kernel density estimate of *yrsedu* is offered in the online appendix.

Unfortunately PROCER does not contain information on income. Instead, a series of dichotomous questions are available inquiring about the presence of services and appliances at the household: (1) land line, (2) paid tv service (Sky, Dish or cable), (3) internet, (4) DVD or Blue Ray, (5) blender, (6) toaster, (7) microwave, (8) fridge, (9) gas or electric stove, (10) washing machine, (11)

[5]The main education levels in Mexico are: None, primary, secondary, preparatory, and first degree.

electric iron, (12) swing machine, (13) fan, (14) video games console, (15) electronic tablet, (16) computer, (17) printer. Adding up these responses we build a variable (*iincome*) that ranges from 0 to 17, with mean 7.6 ($SD = 3.7$) that proxies household permanent income or wealth. The distribution of *iincome* is single picked and slightly right biased. A kernel density estimate of *iincome* is offered in the online appendix. We treat *iincome* as continuous.

Number of bulbs (*bulbs*) in the household is another, different, proxy for permanent income (wealth). The intuition of this proxy is that whenever a household has higher income it tends to have a dwelling with more rooms, more bulbs, and consume more energy. *bulbs* ranges from 0 to 30 and has mean 6.8 ($SD = 3.8$). Its distribution looks similar to a typical income distribution: skewed to the right with a long tail. A kernel density estimate of *bulbs* is offered in the online appendix. We treat *bulbs* as continuous.

lifesat measures life satisfaction in a scale from 1 to 10 and uses the typical question and scale.[6] As it is well know, for a mid income developing country with high inequality, Mexicans tend to score high on life satisfaction and happiness—comparable with USA, Canada, and many high income countries in Europe. Following the known stylized facts, *lifesat* in our sample has mean 8.34 ($SD = 1.4$). Hence, subjects are solidly happy with their life in our analytical sample. Again, we treat this variable as continuous and details on its distribution are offered in the online appendix.

Finally, *hthstat* measures health satisfaction in a 5-item Likert scale. The question is very similar to the typical health self-assessment question.[7] For analysis we inverse coded the variable, so that bad health status has value 1 and very good health status has value 5, and joined categories bad and very bad to avoid small cell size. *hthstat* has mean 3.8 ($SD = 0.59$). To simplify the analysis, we treat this variable as continuous—though, for completeness, we also perform ordered probit analysis. Details for the distribution of *hthstat* are offered in the online appendix.

3.2. Linguistic family heritage

663 (51%) individuals in the analytical sample speak Spanish as an additional language (*SAL*=1) (i.e. they speak Spanish as well as an indigenous language), whereas 637 (49%) are Spanish monolinguals. This is the independent (control) variable of interest. Table 2 offers details about the

[6] The exact phrasing is: In an scale from 0 to 10, where donde 0 is nothing and 10 is all, in general how satisfied are you with your life?

[7] The exact wording is: How do you consider your current health status? Answer options: (1) very good, (2) good, (3) regular, (4) bad, and (5) very bad.

sample distribution of individuals' indigenous language status according to whether their mother and/or father were indigenous language speakers (*inlang*=1). Approximately half of the sample have a mother (father) who speaks or spoke an indigenous language. Importantly, Table 2 shows that mother's (or father's) indigenous language status is a strong predictor of an individual's own indigenous language, and hence SAL, status. There is, however, sufficient variation. For instance, 71% of those whose mother speaks/spoke an indigenous language speak the language themselves. This implies that language is strongly transmitted from mothers to children. But, however strong a predictor it is, an individual still has a good 29% chance of not inheriting his/her mother's indigenous language. Similar stylized facts about language transmission from fathers to children are reported in Table 2.

Table 3 offers details about how much variation there is among mother's and father's indigenous language status in our analytical sample. When a father speaks an indigenous language there is a 13% chance that the mother does not. Similarly, when a father does not speak an indigenous language, there is 23% chance that the mother does. In 88% of the cases both mother and father speak an indigenous language. As these figures show, variation in linguistic family heritage comes from both family sides. Further, though highly correlated, each heritage line has sufficient self-variation.

3.3. Other controls

Unfortunately we do have available a rural/urban indicator. We have controls for sex, education, self-reported skin tone (PRODER), and access to health services. For the family of origin, besides mother's and father's indigenous language status, we have information about both parents' education, occupation, and a proxy for permanent income of the origin family. of Table 1 offers descriptive statistics by SAL status. There some differences in the observable characteristics between the $SAL = 0$ and the $SAL = 1$ samples, starting by the fact that the $SAL = 0$ sample has a higher proportion of females (56%) than the $SAL = 0$ sample (49%). The difference is statistically significant at 1%.

SAL individuals are on average 3.6 months older (insignificant at 5%) and completed two and a half school years less than Spanish monolinguals (significant at 5%). In fact, SAL individuals mostly complete primary school while Spanish monolinguals mostly complete secondary school. There are significant differences across the whole education distribution. At the bottom of the distribution the highest qualification is primary for 41% of the $SAL = 1$ sample, whereas only 29% of the $SAL = 0$ sample report primary as their highest qualification. At the top of the education

distribution only 3.8% of the $SAL = 1$ sample completes a first degree, while 10.5% of the $SAL = 0$ does. Details about the distribution of years of education are given in the online appendix.

Studying the relationship between language and wellbeing is complicated by the fact that language and ethnicity are intimately intertwined. It is important, therefore, to have a good control for ethnoracial differences across study groups. Otherwise the language and ethnoracial effects are likely to be confounded. Given the small size of each ethnic group in Mexico, and in most Latin American countries, it is impossible to control for specific ethnicity in survey based studies as there is little hope to achieve large enough cell sizes to secure valid inference. Using Census data such an approach may be taken. But most Censuses do not collect sufficient details on language heritage and other control variables. Besides, specific ethnicity is hardly the correct control because unlike the USA and Europe, in Latin America "race" and "ethnicity" are somewhat ambiguous categories. Even among minority indigenous groups, the majority of people has some degree of mixed racial background due to the *meztizaje* process (mixing) that started since colonial times and continues today. Main mixtures include white and indigenous, white and black, and white, indigenous, and black. In this context, it is well known in the literature that society strongly stratifies over skin color and class (i.e. income and economic status), rather than over race and specific ethnic group. This is why the Project on Ethnicity and Race in Latin America (PERLA) created the PERLA Color Palette for measuring skin colour tone (see, for instance Telles et al. 2015, Telles and Paschel 2014, Flores and Telles 2012). PERLA has been used in Mexico's National Discrimination Survey but has been never been validated for the Mexican population. Looking to develop a Mexico specific skin color palette, PRODER designed a new 11-scale color palette based on the official PANTONE skin-tone guide. In the present study we use, specifically, subjects' self-assessment of their skin tone. We believe this is the best way to control for "ethnicity" in the Mexican context.

In our sample there are differences on skin tone between SAL and non-SAL individuals. In the online appendix we offer detail showing that individuals in the $SAL = 1$ sample self-classify more frequently into darker categories (A and B) and less frequently in the lighter categories (G, H and I).

We have information about access to health services. In our regressions, however, we do not control for access to health services as this is a post-treatment variable that can be affected by the SAL status.

Regarding the family of origin, we know mother's and father's education. As Table 1 shows, there are differences between the $SAL = 1$ and the $SAL = 0$ samples. In fact, mother's years of education is about 1 year lower for $SAL = 1$ individuals. Differences occur across the whole education distribution. At the bottom, 53% of the $SAL = 1$ mothers have no formal education. In

contrast, only 33% of the $SAL = 0$ mothers did not received any formal schooling. At the top, less than 3% of the $SAL = 1$ mothers completed secondary school, whereas only about 8% of the $SAL = 0$ mothers did. A kernel density estimate of the distribution of mothers' years of education is offered in the online appendix. Similar conclusions can be drawn for fathers' years of education.

Also, from the family of origin, we know mother's and father's occupation. Again, there some differences in the distribution of mother's and father's occupation between the sample who speaks an indigenous languages and the sample that does not. The most important one is mother's labour market participation. Among the $SAL = 1$ sample 73% of the mothers do not work, whereas among the $SAL = 0$ sample the corresponding figure is only 66%. Also, the percentage of self-employed fathers is higher among in the $SAL = 1$ sample (68%) than in the the $SAL = 0$ sample (54%).

Finally we have a proxy for the permanent income, or wealth, of the origin family (short definition). To avoid cluttering we do not disscuss details here on how this variable is build from the dataset. However, such details are offered in the online appendix. The imputed permanent income is a variable that ranges from 0 to 3 with mean 1.68 ($SD = 1.08$). There are some mean differences across language groups. While in the $SAL = 1$ sample imputed permanent income of the origin family (short version) has mean 1.45, in the $SAL = 0$ sample the mean is 1.91. The difference is statistically significant at 1%. A kernel density estimate of the distribution of this variable is offered in the online appendix.

4. Econometric methods

4.1. Exogenous vs endogenous bilingualism

Why would someone belonging to an indigenous ethno-linguistic minority learn to speak Spanish in Mexico? There are various potential reasons. Speaking the language of the majority is important to communicate outside the community for trading at the marketplace and/or the workplace. Interact with government institutions such as schools, health services, or courts outside the community is also important. There is also the fact that some people learn a language (or two or three) from their parents from infancy—i.e. they are born with it and are native speakers—and some learn it later on as an additional language. Therefore, an individual may be monolingual, bilingual, or multilingual by birth. Or she/he may be monolingual by birth and then become bilingual or multilingual as an adult.

While bilingualism (Spanish and indigenous language) by birth passes from parents to children at early stages of life and it may occur fundamentally for 'identity' rather than 'pecuniary' reasons, an individual who becomes bilingual as an adult would normally learn the additional language for

pecuniary motives. In the former case, when one is born bilingual, bilingualism is likely to be exogenous to choices that an individual takes on later in life; say, education choices, labor market participation, occupation, or consuming health damaging substances such as tobacco or alcohol. In the latter case, when one becomes bilingual as an adult, bilingualism is likely to be endogenous because it is quite possible that the individual decides to learn the additional language precisely to improve her/his life chances and wellbeing.

As far as we know, there are no detailed studies and/or statistics about bilingualism among Mexico's indigenous ethnic minorities that may allow researchers to gauge what proportion of bilinguals are born as bilinguals and what proportion become bilinguals as adults. All we know from the Censuses is that among the people who self-declare to be indigenous by culture (34%) are bilinguals, (58%) are Spanish language monolinguals, and (8%) are indigenous language monolinguals. Given the lack of previous knowledge, and while we accumulate sufficient understanding of the topic, from our point of view, researchers studying the relationship between indigenous language and wellbeing among the ethnic minorities of Mexico should consider bilingualism as potentially endogenous.

4.2. Identification strategy

We aim to identify the causal effect of speaking *Spanish as additional language* (SAL) (the *treatment* hereafter) on education, labor, and wellbeing outcomes. As a working hypothesis we will consider treatment to be endogenous. We will use an instrumental variables strategy. Further details on the estimation strategy are given in section 4.3. Here we make explicit our identification strategy.

PRODER 2019 contains information about linguistic family heritage along with self-reported ethnicity. This gives us the opportunity of using parents' language skills as instruments for an individual's own language skills on regressions of our outcome variables on a (binary) SAL. To achieve a control group as clean as possible only individuals who self-declare to be indigenous by culture are included in the analytical sample.

To secure identification of the parameter treatment effect of interest our instrumental variables must comply with three assumptions: (a) relevance, (b) exclusion restriction, and (c) unconfoundness. Together, exclusion restriction and unconfoundness assumptions, ensure that an instrument is exogenous and therefore valid. Conditions (a)-(c) are enough to identify the average treatment effect (ATE) if the treatment effect is homogeneous across the whole population. In such a case the local treatment effect (LATE) and the average treatment effect are equal (ATE=LATE). If there

are heterogeneous effetcs, however, a monotonicity condition (d) is needed to identify the LATE (Angrist and Imbens 1991). Further, adding a conditional independence assumption, the average treatment effect (ATE) and the average treatment effect on the treated (ATET) are identified—more of this in section 4.3. A stable unit treatment assumption (SUTVA) is always needed.

The relevance assumption requires the instrument to be partially correlated with the endogenous variable *SAL*, net of exogenous controls. In our study, the general idea is that conditional on an individual's characteristics, mother's and/or father's ability to speak an indigenous language increase the probability she will speak the indigenous language herself and, therefore, the probability of being SAL. We expect these two partial correlations to be strong. This assumption is testable and we do so in section 5.

The exclusion restriction assumption requires the instrument not to belong to the main regression model. This translates in our study to suppose that, conditional on an individual's characteristics, mother's and/or father's indigenous language status should not affect directly their children's adult wellbeing outcomes. Any potential effect must be indirect through the SAL channel. This our main identification assumption. The direct channel from instruments to endogenous variable is clear as the relationship between parents' and children language skills is expected to be causal, not a simple correlation: Language is transmitted from parents to children from early infancy. Reverse causality is impossible.

It is also clear that what it matters for an individual education, labor, and wellbeing outcomes, is her *personal* ability to speak Spanish; not whether her mother or father spoke the language. Language is a skill that can only deliver returns (or penalties) when an individual uses it in her interaction with others. It helps, for instance, in real/time communication between a buyer and seller to close a transaction at the marketplace. To carry value, such communication must occur at a particular time, with a particular aim, and within a particular context. Reaping the benefits requires produce language within this window of opportunity, it does not matter where, when, and from whom, the individual learned the language. What markets price is the use of the skill at the right time and context—though it can be valued without market. Following these lines of thought we do not expect mother's and father's indigenous language status to affect directly the education, labor, and well being outcomes, of their children. Hence, we believe it is theoretically justified to suppose that parents' language skills do not truly belong to our main response regressions models.

We need now to consider the unconfoundness assumption, which requires the the absence of any unobservables that may affect outcome and instruments at the same time.[8] To gauge this issue

[8]Between instrument and endogenous variable only partial correlation is needed; not necessarily a causal relationship. The existence of a true causal effect between instruments and endogenous variable, however, adds credibility to

we need to ask: Does it any indirect effect from mother's and/or father's indigenous language status to an individual's wellbeing (education / labor) outcomes comes really only through the language channel?

It is here, on unconfoundness, where our identification strategy faces its strongest challenges. From an start, parents' language should affect their own education, labor market, and wellbeing outcomes. And, while parents' ability to speak an indigenous language is not expected to have a direct effect on their children's wellbeing (education / labor), their education and income will probably do. Hence, effectively, this is a indirect effect that does not goes through the language channel. We believe this is the main threat to our identification strategy. Fortunately, PRODER was explicitly designed to study how racism and ethnoracial perceptions of discrimination relate to socioeconomic inequality in Mexico, and put effort on measuring how much inequality is transmitted from parents to children—i.e. PRODER had particular interest on studying inter-generation inequality transmission. This means that PRODER collected detailed information about mother's and father's characteristics. In particular, we have detailed information on mother's and father's education, occupation, and a proxy for permanent income of the household of origin at the time our subjects were aged 14. After conditioning on these variables, we believe, there are no other obviously important variables left in the error terms that could substantially challenge the unconfoundness assumption.

Summarizing: We sustain that all conditions for having valid instrumental variables for our study hold and that our identification strategy is sound.

The monotonicity assumption calls for the instrument never to reduce the probability of treatment. Because we have no indigenous language monolinguals in our sample, this translates in our application to requiring that an individual's probability of speaking an indigenous language—and so to be SAL—will never decrease by her mother or father speaking the language. This sounds quite reasonable and find no theory reasons to argue otherwise. Under monotonicity we can exclude the existence of defiers and ensures identification of the LATE even when the treatment effect is not homogeneous across the population (Angrist and Imbens 1991).

Finally, we need to exclude the possibility that treatment of one individual spills over other individuals. This is known as the stable unit treatment assumption (SUTVA). In our study, SUTVA requires that, outside family members, the fact that one individual speaks an indigenous language does not change the probability that other members in the community would speak the language as well. This assumption may be invalid if the ethno-lingustic community to which the individual

our identification strategy as one can control for all available potential observable counfounders and be sure that some real and strong relationship between instrument and endogenous variable should remain.

belongs is too small. A few houses in a small villa, for instance. In contrast, if the ethno-lingustic community to which the individual belongs is large enough, it is safe to assume one particular individual cannot substantially affect the probability that other, non family related, individuals in her community will speak the indigenous language. Because PRODER oversamples areas of Mexico where relative large ethno-lingustic communities live, we believe that SUTVA reasonably holds in our study population.

4.3. Estimation strategy

Our aim is to estimate the causal effect of speaking Spanish as additional language (the 'treatment') $t = 1$ on a series of wellbeing outcomes y. We say that individuals who do not speak Spanish as additional language are untreated $t = 0$ units. We consider, one-by-one and separately, seven different responses: Current work status *work* (binary), years of education *yrsedu* (continuous), imputed permanent income *iincome* (continuous), number of bulbs in the household *bulbs* (continuous), life satisfaction $lifesat$ (continuous), and health status $hthstat$ (ordinal).

The parameters of interest are: (1) The local treatment effect (LATE), (2) the average treatment effect (ATE), and/or (3) the average treatment effect on the treated (ATET). As we will see shortly, depending on the model, in some cases the three quantities are the same, while in other cases the three quantities are different. Not all estimation strategies deliver an estimate for the three quantities.

As a working hypothesis we consider treatment t as being potentially endogenous in a regression of y on t and control variables $\mathbf{x}$. To address the endogeneity of t we use father's and mother's indigenous language status—both binary—as instruments (or to specify exclusion restrictions) for t. Two different models are estimated:

(a) *Homogeneous treatment effect* model (HoTEM). A three-equation system composed by one equation for the main response y and two equations that determine the endogenous treatment t. The effect of t on y is constant across the population and does not depend on individual characteristics $\mathbf{x}$. Hence, in this model LATE = ATE = ATET.

(b) *Heterogeneous treatment effects* model (HeTEM). A four-equation system that explicitly implements a potential outcomes model (PO) for y, composed by one equation for the main response when a unit is treated y_1, one equation for the main response when a unit untreated y_0, and two equations that determine treatment status—and, therefore, whether y_0 or y_1 is observed. The effect of treatment is different for each individual in the population depending

upon $\mathbf{x}$. Hence, in general, LATE $\neq$ ATE $\neq$ ATET.

For each model, HoTEM and HeTEM, we implement a set of different estimators. Each estimator is consistent for the parameter(s) of interests under different conditions. Details are discussed in the the next two subsections following notation from `etregress` and `eteffects` (StataCorp 2025).

4.3.1. Homogeneous effect

The structural system of interest is,

$$y_i = \mathbf{x}_i\beta + \theta t_i + a_s + u_i, \tag{1}$$

$$t_i^* = \mathbf{w}_i\gamma + b_s + v_i, \tag{2}$$

$$t_i = \mathbb{1}(t_i^* > 0) \tag{3}$$

where y_i is the i-th individual's response (outcome), which is always observed, t_i is the treatment status (In our case SAL), and $\mathbb{1}(\cdot)$ is the indicator function. t_i^* is a latent (unobserved) variable that can be interpreted as the utility of treatment. Treatment occurs only when the utility of treatment crosses certain threshold; here normalized to zero without loss of generality. Equation (2) and (3) define the data generation mechanism of the, always observed, treatment status t_i. $\mathbf{x}$ and $\mathbf{w}$ are vectors of exogenous explanatory variables (including the constant term), while β and γ represent conformable vectors of coefficients. $\mathbf{x}$ and $\mathbf{w}$ may contain common variables. However, we assume that there is at least one variable in $\mathbf{w}$ not present in $\mathbf{x}$. So, the system is identified on the basis of first moment restrictions. Next, a_s and b_s are a set of state-level fixed-effects, whereas u and v are error terms. We suspect $E(u|v,\mathbf{q}) \neq 0$; with $\mathbf{q}$ representing the whole set of exogenous variables in the system. As a consequence, we say that the treatment t is endogenous. The coefficient of interest is θ. This is the causal effect of SAL on the main response y. In this model the effect of treatment is constant across the whole population. As a consequence, a consistent estimator for θ is also a consistent estimator for the LATE=ATE=ATET.

We implement four different estimators: (a) Ordinary least squares (OLS), (b) Two-stage least squares (2SLS), (c) Maximun likelihood (MLE), and (d) Control function (CF).

The OLS estimator is inconsistent for θ unless, against our working hypothesis, $E(u|v,\mathbf{q}) = 0$ is true. In such a case t is exogenous and the main response equation (1) can be fitted without explicitly modeling the treatment data generating mechanism. Conditioning on $\mathbf{x}$ and t suffices to obtain a consistent estimator for θ.

If $E(u|v,\mathbf{q}) \neq 0$, as we suspect, then treatment endogeneity needs to be explicitly addressed. A 2SLS estimator is fitted using father's and mother's indigenous language status as instruments for

SAL. 2SLS delivers a consistent estimator of θ if conditions (a)-(c) on section 4.2 hold. We do not give more detail as the 2SLS is a well known and popular estimator in the econometrics literature.

A second alternative is to assume that (u,v) are bivariate normal and implement a maximum likelihood estimator. The contribution of the i-th individual to the log likelihood is,

$$\begin{aligned}\log L_i = (1-t_i)\left[\log\Phi\left(\frac{-\mathbf{w}_i\gamma-\left(\frac{\rho}{\sigma}\right)(y_i-\mathbf{x}_i\beta)}{\sqrt{1-\rho^2}}\right)-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\beta}{\sigma}\right)^2-\log\left(\sqrt{2\pi\sigma^2}\right)\right]+ \\ t_i\left[\log\Phi\left(\frac{\mathbf{w}_i\gamma+\left(\frac{\rho}{\sigma}\right)(y_i-\mathbf{x}_i\beta-\theta)}{\sqrt{1-\rho^2}}\right)-\frac{1}{2}\left(\frac{y_i-\mathbf{x}_i\beta-\theta}{\sigma}\right)^2-\log\left(\sqrt{2\pi\sigma^2}\right)\right].\end{aligned} \tag{4}$$

Where $\Phi(\cdot)$ is the commulative normal distribution, $\mathrm{Var}(u)=\sigma^2$, and $\rho=\mathrm{Cor}(u,v)$. The log-likelihood is maximized using a Newton-Ramphson algorithm. At the global maximum $-\widehat{H}^{-1}$ provides an estimator for the covariance matrix; with $\widehat{H}$ representing the fitted Hessian. Heteroskedasticity robust standard errors can be obtained using the sandwich estimator of the covariance matrix (White 1980, Huber et al. 1967). The ML estimator is a consistent and fully efficient estimator for θ. Small departures from normality, however, renders the ML estimator inconsistent.

A third alternative is to implement a control function estimator (CF) estimator, which is more robust to misspecification of the (u,v) distribution. Under bivariate normality $E(u|v,t,\mathbf{q})=\rho\sigma h$, with $h(\cdot)$ representing the hazard function

$$h = t\left[\frac{\phi(\mathbf{w}\gamma+b)}{\Phi(\mathbf{w}\gamma+b)}\right]+(1-t)\left[\frac{-\phi(\mathbf{w}\gamma+b)}{1-\Phi(\mathbf{w}\gamma+b)}\right] \tag{5}$$

and $\phi(\cdot)$ the standard normal density. It can be shown that

$$E(y|t,\mathbf{q}) = \mathbf{x}\beta+\theta t+\rho\sigma h+a_s, \tag{6}$$

$$\mathrm{Var}(y|t,\mathbf{q}) = \sigma^2\left\{1-\rho^2\left[h(h+\mathbf{w}\gamma)\right]\right\}. \tag{7}$$

Hence, it is possible then to augment equation (1) to get

$$y_i = \mathbf{x}_i\beta+\theta t_i+\rho\sigma h_i+a_s+\varepsilon_i, \tag{8}$$

with $\varepsilon = u-E(u|v,t,\mathbf{q})$, so that $E(\varepsilon|v,t,\mathbf{q})=0$ by construction. Therefore, if a consistent estimator for h is available, the control function $\widehat{h}$ deals with the endogeneity of t. This suggests a two-step estimator: (1) fit a probit regression of t on $(\mathbf{w},b)$ to obtain $\widehat{h}$; (2) fit a OLS regression of y on $(\mathbf{x},a,\widehat{h})$ to obtain a consistent estimator for θ. Standard errors must be adjusted for the varia-

tion of the first-stage parameters and a sandwich estimator of the covariance matrix is available to obtain heteroskedasticity robust standard errors.

Instead of fitting a two-step CF estimator, it is possible to implement a one-step GMM CF estimator by means of stacking the moment conditions. Three error functions are involved

$$\begin{aligned}
\varepsilon_t &= t\left[\frac{\phi(\mathbf{w}\gamma+b)}{\Phi(\mathbf{w}\gamma+b)}\right]+(1-t)\left[\frac{-\phi(\mathbf{w}\gamma+b)}{1-\Phi(\mathbf{w}\gamma+b)}\right], && (9)\\
\varepsilon_y &= y-\mathbf{x}_i\beta-\theta t_i-\rho\sigma\varepsilon_t-a_s, && (10)\\
\varepsilon_{var} &= \varepsilon_y^2-\sigma^2\left\{1-\rho^2\left[\varepsilon_t(\varepsilon_t+\mathbf{w}\gamma)\right]\right\}. && (11)
\end{aligned}$$

Let $\mathbf{z}=(\mathbf{x},t,h)$ and write,

$$\mathbf{Z}=\begin{bmatrix}\mathbf{z} & 0 & 0\\ 0 & \mathbf{w} & 0\\ 0 & 0 & 1\end{bmatrix} \quad \text{and} \quad \varepsilon=\begin{bmatrix}\varepsilon_y\\ \varepsilon_t\\ \varepsilon_{var}\end{bmatrix}.$$

Then stack the moment conditions

$$\mathbf{s}=\mathbf{Z}'\varepsilon,$$

and define the one-step GMM estimator for $\pi=(\beta',\theta,\gamma',\rho,\sigma)'$ as the vector $\widehat{\pi}$ that satisfies

$$\frac{1}{N}\sum_i \mathbf{s}_i=\mathbf{0}.$$

A Huber-White estimator of the covariance matrix is estimated forming the usual sandwich formula. In the present analysis we use this one-step CF GMM estimator for analysis.

4.3.2. Heterogeneous effects

To allow heterogeneous effects we generalize system (1)-(3) using a potential outcomes (PO) framework. The structural system is now,

$$\begin{aligned}
y_{0i} &= \mathbf{x}_i\beta_0+a_s+u_{0i}, && (12)\\
y_{1i} &= \mathbf{x}_i\beta_1+a_s+u_{1i}, && (13)\\
t_{si}^* &= \mathbf{w}_i\gamma+b_s+v_i, && (14)\\
t_i &= \mathbb{1}(t_i^*>0) && (15)\\
y_i &= (1-t_i)y_{0i}+t_iy_{1i}. && (16)
\end{aligned}$$

Three different estimators are considered: (i) Maximum likelihood (PO-MLE), (ii) control function under joint normality (PO-CL joint normality), and (iii) control function under the assumption that the conditional mean error is linear (PO-CL under linearity of $E(u|v)$)—which is weaker than calling for full-blown joint normality. As we discussed earlier, control function estimators are more robust to departures from distributional assumptions than ML estimators—which loose consistency with small violations to joint normality.

To implement the maximum likelihood estimator we suppose $(u_0, u_1, v)' \sim MVN(\mathbf{0}, \Sigma)$, with covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_0\rho_0 \\ \sigma_{01} & \sigma_1^2 & \sigma_1\rho_1 \\ \sigma_0\rho_0 & \sigma_1\rho_1 & 1 \end{bmatrix}.$$

σ_{01} is not identified as y_0 and y_1 are never observed simultaneously. The contribution of the i-th individual to the log-likelihood is

$$\begin{aligned} \log L_i = (1-t_i)\left[\log\Phi\left(\frac{-\mathbf{w}_i\gamma - \left(\frac{\rho_0}{\sigma_0}\right)(y_i - \mathbf{x}_i\beta_0)}{\sqrt{1-\rho_0^2}}\right) - \frac{1}{2}\left(\frac{y_i - \mathbf{x}_i\beta_0}{\sigma_0}\right)^2 - \log\left(\sqrt{2\pi\sigma_0^2}\right)\right] + \\ t_i\left[\log\Phi\left(\frac{\mathbf{w}_i\gamma + \left(\frac{\rho_1}{\sigma_1}\right)(y_i - \mathbf{x}_i\beta_1)}{\sqrt{1-\rho_1^2}}\right) - \frac{1}{2}\left(\frac{y_i - \mathbf{x}_i\beta_1}{\sigma_1}\right)^2 - \log\left(\sqrt{2\pi\sigma_1^2}\right)\right]. \end{aligned} \tag{17}$$

The global maximum $\widehat{\pi} = (\widehat{\beta}_0', \widehat{\beta}_1', \widehat{\gamma}', \widehat{\rho}_0, \widehat{\rho}_1, \widehat{\sigma}_0, \widehat{\sigma}_1)'$ is a consistent and fully efficient estimator whenever the joint normality assumption holds.

Maintaining multivariate normality, a control function estimator (PO-CL under joint normality) can be implemented using the fact that

$$E(y|t, \mathbf{q}) = (1-t)\left(\mathbf{x}\beta_0 + \frac{\rho_0}{\sigma_0}\varepsilon_t\right) + t\left(\mathbf{x}\beta_1 + \frac{\rho_1}{\sigma_1}\varepsilon_t\right) \tag{18}$$

$$\mathrm{Var}(y|t=0, \mathbf{q}) = \sigma_0^2\left\{1 - \rho_0^2\left[\varepsilon_t(\varepsilon_t + \mathbf{w}\gamma)\right]\right\} \tag{19}$$

$$\mathrm{Var}(y|t=1, \mathbf{q}) = \sigma_1^2\left\{1 - \rho_1^2\left[\varepsilon_t(\varepsilon_t + \mathbf{w}\gamma)\right]\right\}. \tag{20}$$

Let $\varepsilon_y = y - E(y|t, \mathbf{q})$, $\varepsilon_{var_0} = \varepsilon_y^2 - \mathrm{Var}(y|t=0, \mathbf{q})$, and $\varepsilon_{var_1} = \varepsilon_y^2 - \mathrm{Var}(y|t=1, \mathbf{q})$, with ε_t defined

as in (9). Finally, let $\mathbf{z} = [\mathbf{x}, th, (1-t)h]$ and write

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z} & 0 & 0 & 0 \\ 0 & \mathbf{w} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \varepsilon = \begin{bmatrix} \varepsilon_y \\ \varepsilon_t \\ \varepsilon_{var_0} \\ \varepsilon_{var_1} \end{bmatrix}.$$

Stack the moments $\mathbf{s} = \mathbf{Z}'\varepsilon$ and define the one-step GMM estimator $\widehat{\pi} = (\beta_0', \beta_1', \gamma', \rho_0, \rho_1, \sigma_0, \sigma_1)'$ as the vector that satisfies the sample moment conditions

$$\frac{1}{N}\sum_i \mathbf{s}_i = \mathbf{0}.$$

An assumption of full multivarite normality is stronger than needed. In fact, it suffices to assume that the conditional mean error is linear; which is a weaker condition. Suppose then,

$$E\left(u_j|v\right) = \tau_j v;\ j = 0, 1,$$

which implies,

$$E\left(y|t, \mathbf{q}\right) = (1-t)\left(\mathbf{x}\beta_0 + \tau_0 v\right) + t\left(\mathbf{x}\beta_1 + \tau_1 v\right).$$

Under these assumptions is possible to estimate the ATE and the ATET, together with the potential outcome mean for the non treated (POM_0), without estimating correlations ρ_0 and ρ_1 or variances σ_0^2 and σ_1^2. Define,

$$\varepsilon_y = y - E\left(y|t, \mathbf{q}\right), \tag{21}$$

$$\varepsilon_t = t\left[\frac{\phi\left(\mathbf{w}\gamma + b\right)}{\Phi\left(\mathbf{w}\gamma + b\right)}\right] + (1-t)\left[\frac{-\phi\left(\mathbf{w}\gamma + b\right)}{1 - \Phi\left(\mathbf{w}\gamma + b\right)}\right], \tag{22}$$

$$\varepsilon_{POM0} = \mathbf{x}\beta_0 + \tau_0 v - \text{POM}_0, \tag{23}$$

$$\varepsilon_{\text{ATE}} = \mathbf{x}\beta_0 + \tau_1 v - \text{POM}_0 - \text{ATE}. \tag{24}$$

Stack the moments and obtain the one-step GMM estimator for $\pi = \left(\beta_0', \beta_1', \gamma',\ \text{POM}_0, \text{ATE}\right)'$ as usual. For implementation, a consistent estimator for v is needed. $\widehat{v} = t - \Phi\left(\mathbf{w}\widehat{\gamma} + \widehat{b}_s\right)$, with $\left(\widehat{\gamma}', \widehat{b}_s'\right)'$ fitted by probit will do. To estimate the ATET instead, the last condition is repleced by

$$\varepsilon_{\text{ATET}} = \frac{N}{N_t}\left(\mathbf{x}\beta_0 + \tau_1 v - \text{POM}_0\right) - \text{ATET}$$

where N_t represents the number of treated units. For all estimators just discussed, robust inference can be performed using Huber-White sandwich standard errors.

5. Main results

Table 4 presents results for the whole sample. Each column reports regression for one response variable, which is labeled accordingly. In all models control variables include: sex, age, PRODER skin tone scale, mother's education, father's education, mother's occupation, father's education, origin family imputed permanent income, and state level fixed effects. Robust standard errors are used for inference. To ease interpretation and comparison between models and estimates, table 4 reports only the coefficient on the binary indicator for *SAL*; which is the endogenous treatment of interest. When possible, the estimates for the ATE and the ATET of *SAL* are reported.

5.1. Homogeneous treatment effect

For the model with homogeneous treatment effect four sets of estimates are included: (i) OLS, (ii) 2SLS, (iii) Maximun likelihood (MLE), and (iv) Control function (CF).

We start considering results for current work status (work). This is a binary response. The OLS coefficient estimate is negative -0.02 and insignificant at 5%. This is a consistent estimate for the LATE under the assumption that the treatment is exogenous. Moving to 2SLS we find a positive 0.02 LATE, which is also insignificant at 5%. Our instruments (mother's and father's indigenous language status) are strong predictors of an individual's indigenous status, net of all controls in the system. In fact, the Olea and Pflueger (2013) first stage effective F statistic is 256 (referred as F^{eff} hereafter); which rejects the null of weak instruments at 5%. Results from the 1st stage are presented in the online appendix. A Hansen over identification test reports a $J-stat = 0.48$ ($p-val = 0.49$). Hence, we fail to reject the null that the over identifying restriction is valid at 5%. In other words, there is empirical evidence suggesting that our identification strategy is valid and we are indeed obtaining a consistent estimate of the true causal effect. Finally, a robust endogeneity test fails to reject the null that the treatment is exogenous ($\chi^2(1) = 1.19; p-val = 0.28$). Results for MLE and CF are absent in table 4 because these estimators are not adequate for a binary outcome—as they are models for a continuous response. However, results from nonlinear models are presented in the online appendix showing similar results.

Next, we discuss results for years of education (*yrsedu*). This is a continuous variable. OLS finds a negative marginal effect of *SAL* on *yrsedu* of about -1.11 years, which is significant at

1%. The 2SLS estimate is -0.91 and significant at 5%; equivalent to a drop in education of about -0.2 standard deviations—and hence, economically relevant. Notice that OLS estimate falls within the 95% confidence interval of the 2SLS estimate. So, ignoring covariances, they are statistically the same. In terms of the quality of the 2SLS estimator we find that a $F^{eff} = 256$ rejects the null of weak instruments and an over identification test fails to reject the null $(J-stat = 0.18; p-val = 0.67)$. Hence, we have evidence that our identification strategy works well and we are effectively obtaining a consistent estimate of the true causal effect. Again, we fail to reject the null of exogeneity of the treatment $(\chi^2(1) = 0.30; p-val = 0.58)$. Results from the maximum likelihood and control function estimators are quite similar. Our estimates for the $LATE = ATE = ATET = -0.99$ are consistent with those reported for the 2SLS estimates.

Going from not speaking Spanish as native ($SAL = 0$) to speaking the language as an additional language ($SAL = 1$) reduces imputed permanent income by about -1.16 units in an scale of 17. This an economically relevant average treatment effect; equivalent to a decrement on income of about -0.33 standard deviations. Again, evidence suggest that our identification strategy works well with an $F^{eff} = 256$ and $J-stat = 0.13$ $(p-val = 0.71)$. Moreover, we fail to reject the null of an exogenous treatment across all estimators.

Similar results are found when the response variable is the number of bulbs in the household (*bulbs*), which is also a proxy for permanent income. Our estimate for the LATE=ATE=ATET is -1.46 units, equivalent to decrement of -0.4 standard deviations, and significant at 1% across all alternative estimators. As before, $F^{eff} = 256$ and $J-stat = 0.11$ $(p-val = 0.11)$ suggest instruments are valid and our estimators do not suffer from a weak instruments problem. As a consequence, we are confident that this is a consistent estimate or the true causal effect. While the null of exogeneity cannot be rejected for the 2SLS estimator $(\chi^2(1) = 2.8; p-val = 0.15)$, the null for H_0: $\rho = 0$ is rejected by the MLE $(\chi^2(1) = 4.1; p-val = 0.04)$ and CF $(\chi^2(1) = 4.2; p-val = 0.04)$ estimators.

Moving to the responses for life satisfaction (*lifesat*) and health status (*hthstat*) our results suggests that instruments are valid and not weak in both cases. However, our estimate of the ATE is insignificant at 5% across all estimators for both responses. That is, our results suggest that speaking Spanish as an additional language has zero effect on *lifesat* and *hthstat*. Treatment exogeneity cannot be rejected across all estimators.

5.2. Heterogeneous treatment effects

For the model with heterogeneous treatment effects (potential outcomes) three sets of estimates are included: (i) PO-MLE, (ii) PO-CF under joint normality, (iii) PO-CF under linearity of $E(u|v)$. Here, in general, LATE$\neq$ATE$\neq$ATET. The last three panels of table 4 report results for the potential outcome models.

We find an statistically insignificant (zero or null) treatment effect of SAL on *lifesat* and *hthstat* across all alternative estimators. So, we no further discuss those responses.

For years of education (*yrsedu*) we find an estimate of the $ATE = -1.13$ (significant at 1%) and the $ATET = -0.83$ (significant at 10%) from the Maximum likelihood estimator. Control function under normality gives similar estimates: $ATE = -1.13$ and $ATET = -0.87$. Finally, control function under linearity of $E(u|v)$ gives estimates $ATE = -1.27$ and $ATET = -0.99$. In the latter case both ATE and ATET are statistically significant at 5%. These effects represent a drop in education equivalent to -0.28 and -0.21 standard deviations, which are economically relevant. All estimators fail to reject the null of an exogenous *SAL* treatment. In particular, the PO-CF under linearity of $E(u|v)$ fails to reject the null of exogeneity with $\chi^2(1) = 0.69$ and $p - val = 0.71$.

Moving to imputed permanent income (*iincome*) we find $ATE = -1.14$ (significant at 1%) and $ATET = -0.89$ (significant at 1%) for the PO-MLE estimator. Control function estimates are quite similar. In fact, under under linearity of $E(u|v)$ we find $ATE = -1.15$ (significant at 1%) and $ATET = -1.76$ (significant at 1%). These effects are equivalent to a drop in income of -0.33 and -0.51 standard deviations, respectively. A test for endogeneity of treatment fails to reject the null ($\chi^2(1) = 0.01$; $p - val = 0.93$). So, empirical evidence indicate that *SAL* is in fact an exogenous treatment.

Finally, for (*bulbs*) the PO-CF under linearity of $E(u|v)$ reports estimates $ATE = -1.46$ (significant at 1%) and $ATET = -1.79$ (significant at 1%), respectively. These effects are equivalent to a drop in income of -0.40 and -0.49 standard deviations. In this case we find weak evidence of endogeneity as the test is able to reject the null with a $\chi^2(1) = 5.39$ and $p - val = 0.07$.

6. Results by gender

Tables 5 and 6 present results from linear models for females and males, respectively. To avoid cluttering, we do not discuss in detail results from each model and estimators fitted separately on the females and males samples. Instead, we give a short summary of the findings.

The sample size is $N = 681$ (52% of the whole sample) for females and $N = 611$ for males. Looking at the 2SLS estimator, we find that the effective first-stage F^{eff} is 118 for females and

141.27. Hence, for both genders, 2SLS hardly suffers from a problem of weak instruments. Also, a Hansen over identification test fails to reject the null for all responses and for both sexes at 1% across all response variables. Hence, we have strong evidence that our 2SLS estimator is well identified for both sexes and across all responses.

An endogeneity test for the SAL treatment fails to reject the null for both males and females across all models and estimators at 5%. So, treatment is exogenous for both females and males. Consistent with results for the whole sample, for both females and males, we find that the SAL has a null effect on current work status, life satisfaction, and health status. This is true across all models and estimators—including nonlinear models.

We jump now to discuss results from our best preferred specification, which is the potential outcomes control (PO-CF) function estimator under the assumption of linearity of $E(u|v)$. As we mentioned before, this estimator allows for heterogeneous treatment effects and falls short of requiring multivariate normality.

We find a point estimate of the ATE on years of education of -1.49 (significant at 10%) for females and -1.49 (significant at 5%) for males. The ATET is insignificant in both cases. Clearly, splitting the sample has importantly affected the precision. In fact, with respect to the whole sample, SEs almost doubled for females and increased by a factor of 0.16 for males. Notice, however, ignoring covariances, that the ATE point estimates from PO-CF under linearity of $E(u|v)$ fall well within the 95% confidence interval of the OLS estimate of the LATE for both sexes, and that the OLS estimate is statistically significant at 5% whilst treatment is exogenous. Putting all the evidence together we can say we find a significant at 5% ATE of *SAL* on *yrsedu* of about -1.5 years and no much evidence of treatment heterogeneity across sexes.

Moving to imputed permanent income point estimates of the ATE from PO-CF under linearity of $E(u|v)$ is -1.36 (significant at 5%) for females and -1.17 (significant at 1%) for males. So, there is some evidence of gender treatment heterogeneity. The reader should be aware, however, that, ignoring covariances, the 95% confidence intervals overlap. Hence, the ATEs for males and females are hardly statistically different. Again, point estimates for the ATE from PO-CF under linearity of $E(u|v)$ fall well within the 95% confidence interval of the OLS estimates. Similar conclusions can be drawn from the ATET estimates, which are significant at 1% for both males and females. Notice that 95% confidence intervals for ATET and ATE overlap.

Finally for *bulbs* estimates of the ATE from PO-CF under linearity of $E(u|v)$ is -2.27 (significant at 1%) for females and -0.89 (insignificant at 5%) for males. The point estimate for males falls outside the 95% confidence interval of the females estimate. Hence, in this case there is indeed evidence of treatment heterogeneity. Notice that, among all response variables, *bulbs* is the

only case in which we find evidence that *SAL* is indeed endogenous.

7. Robustness checks

For completeness the online appendix presents results from fitting an endogenous treatment logit regression for current work status. Similarly, regression results are presented from an endogenous treatment ordinal probit for health status. Findings indicate that in both cases the SAL treatment is exogenous and has a null effect at any conventional significance level.

Also, in the online appendix we offer results from instrumental variables quantile regression for years of education, imputed permanent income, bulbs, and life satisfaction—all the continuous variables considered in the present study. Quantiles Q_{25}, Q_{50}, and Q_{75} are considered. Here coefficients on SAL have a LATE interpretation. While results from education and permanent income suggest that SAL treatment effects are larger at the top of the conditional distribution, diagnostics fail to reject the null of a constant effect in both cases—So a QR is not justified. In both cases, as well, a test for the exogeneity of the treatment fails to reject the null. Similar results are found for the number of bulbs at the household. For life satisfaction, we find point estimates for the LATE that are statistically insignificant at all considered quantiles. In fact, our QR point estimates fall with the 95% confidence interval of the OLS point estimate. Again, diagnostic tests fail to reject the null of a constant effect as well as the null of exogenous treatment.

8. Discussion and conclusions

We estimate the causal effect of speaking *Spanish as an additional language* (SAL)—as opposed to speaking it as a native—on education, labor, and wellbeing outcomes, among the indigenous ethno-lingustic minorities of Mexico using data that are particularly suited to study the topic: The Ethno-Racial Discrimination in Mexico Project Survey 2019 (PRODER). To secure a suitable control group only data from individuals who self-declared to be indigenous are analyzed. The sample does not contain indigenous language monolinguals—hence, all those who speak an indigenous language are SAL. Potential treatment endogeneity is addressed using mother's and father's indigenous language status as instrument for individuals' own language status. Models allowing for homogeneous and heterogeneous treatment effects are fit with 2SLS, Maximum likelihood, and control function estimators. Instruments are strong predictors of treatment status and over identification tests show empirical evidence that support the claim that they are valid. Unlike previous work, we are able to control for potential key confounders that are rarely observed, including

parental education and occupation, as well as family imputed income (wealth). This gives us confidence that we are estimating true causal effects.

Findings indicate that indigenous individuals in Mexico who speak Spanish as an additional language (SAL) complete, on average, one year less of education than Spanish monolinguals individuals. This is a reduction of approximately -0.2 standard deviations in educational attainment. Furthermore, speaking Spanish as a second language is associated with a penalty of -23% in a measure of permanent income (or wealth). Regarding another proxy of wealth—number of light bulbs in the household——we find a SAL treatment effect of -30%. Our income/wage treatment effect has the same sign and is of similar in size to additional language (AL) estimates reported previously in the literature (see particularly estimates by Grenier 1984, Gonzalez 2005, Chiswick and Miller 2016, Miranda and Zhu 2013). No statistically SAL effects are found on current employment status, life satisfaction, or health status; nor evidence of heterogeneous treatment across genders.

Overwhelming evidence shows that SAL is, in fact, exogenous for education and permanent income. This result, however, could not be known had we not explicitly modeled treatment endogeneity.

While we are confident of identifying true SAL causal effects, further research is needed to investigate what are the exact mechanisms behind the negative effects on education and permanent income we find. Various mechanisms may be at play: (a) low Spanish proficiency hindering communication, (b) occupation segregation preventing SAL individuals access to higher paid jobs, (c) accent and non-native speech pattern discrimination, (d) residential segregation, (e) differential migration, and/or (f) differential marriage patterns. Unfortunately, PRODER does not contain either objective nor subjective measures of Spanish proficiency. Nor we have information that may help us to study the role of accent and non-native speech pattern discrimination, residential segregation, or migration. Simple tabulations show that 63% of *SAL* individuals have partners who speak an indigenous language, whereas only 7% of the non-*SAL* individuals do. Hence, differential marriage patterns by language group can play a role explaining permanent income differentials. Tabulations also show that some degree of occupation segregation by SAL status exists. 23% (53%) of the *SAL* sample are private employees (self-employees). In contrast, for the non-*SAL* sample the corresponding figure is 38% (42%). We do not have, however, finer occupation categories, sufficient sample size, or way of investigating what role language plays in each category, to attempt investigating how much of our estimated treatment effects run trough occupation segregation by language groups. All these questions stay open for future research.

In terms of public policy the present work has four main implications: (1) Speaking Spanish as

additional language among the indigenous ethno-linguistic minorities of Mexico carries important negative returns on education and permanent income that require further attention as a source of social and economic disadvantage; (2) the effect is causal rather than just a correlation; (3) better data and further research are needed to understand the underlying mechanisms; finally, (4) implementing actions of public policy are needed to address/reduce the language education and income gap created by SAL among the indigenous ethno-linguistic minorities of Mexico.

References

Aguilar-Rodriguez, A., Miranda, A., and Zhu, Y. (2018). Decomposing the language pay gap among the indigenous ethnic minorities of mexico: Is it all down to observables? *Economics Bulletin*, 38(2):689–695.

Angrist, J. and Imbens, G. (1991). Sources of identifying information in evaluation models.

Berman, E., Lang, K., and Siniver, E. (2003). Language-skill complementarity: returns to immigrant language acquisition. *Labour Economics*, 10(3):265–290.

Bleakley, H. and Chin, A. (2004). Language skills and earnings: Evidence from childhood immigrants. *Review of Economics and statistics*, 86(2):481–496.

Bloomfield, L. (1933). *Language*. Henry Holt.

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

Chiswick, B. R. (2009). The economics of language for immigrants: An introduction and overview. *The education of language minority immigrants in the United States*, 3568:72–91.

Chiswick, B. R. and Miller, P. W. (1995). The endogeneity between language and earnings: International analyses. *Journal of labor economics*, 13(2):246–288.

Chiswick, B. R. and Miller, P. W. (2015). International migration and the economics of language. In *Handbook of the economics of international migration*, volume 1, pages 211–269. Elsevier.

Chiswick, B. R. and Miller, P. W. (2016). Does bilingualism among the native born pay?

Chiswick, B. R. and Miller, P. W. (2018). Do native-born bilinguals in the us earn more? *Review of Economics of the Household*, 16:563–583.

Chiswick, B. R., Patrinos, H. A., and Hurst, M. E. (2000b). Indigenous language skills and the labor market in a developing economy: Bolivia. *Economic development and cultural change*, 48(2):349–367.

De la Fuente Stevens, D. and Pelkonen, P. (2023). Economics of minority groups: Labour-market returns and transmission of indigenous languages in mexico. *World Development*, 162:106096.

Donado, A. (2017). Foreign languages and their impact on unemployment. *Labour*, 31(3):265–287.

Dustmann, C. and Fabbri, F. (2003). Language proficiency and labour market performance of immigrants in the uk. *The economic journal*, 113(489):695–717.

Dustmann, C. and Soest, A. v. (2001). Language fluency and earnings: Estimation with misclassified language indicators. *Review of Economics and Statistics*, 83(4):663–674.

Dustmann, C. and Van Soest, A. (2002). Language and the earnings of immigrants. *ILR Review*, 55(3):473–492.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of phonetics*, 15(1):47–65.

Flores, R. and Telles, E. (2012). Social stratification in mexico: Disentangling color, ethnicity, and class. *American sociological review*, 77(3):486–494.

Gao, W. and Smyth, R. (2011). Economic returns to speaking 'standard mandarin'among migrants in china's urban labour market. *Economics of Education Review*, 30(2):342–352.

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

Godoy, R., Reyes-García, V., Seyfried, C., Huanca, T., Leonard, W. R., McDade, T., Tanner, S., and Vadez, V. (2007b). Language skills and earnings: Evidence from a pre-industrial economy in the bolivian amazon. *Economics of Education Review*, 26(3):349–360.

Gonzalez, L. (2005). Nonparametric bounds on the returns to language skills. *Journal of Applied Econometrics*, 20(6):771–795.

Grenier, G. (1984). The effects of language characteristics on the wages of hispanic-american males. *Journal of Human Resources*, pages 35–52.

Grenier, G. (1987). Earnings by language group in quebec in 1980 and emigration from quebec between 1976 and 1981. *Canadian Journal of Economics*, pages 774–791.

Grenier, G. Z. et al. (2021). The value of language skills. *IZA World of Labor*.

Grogger, J. (2019). Speech and wages. *Journal of Human Resources*, 54(4):926–952.

Grosjean, F. (2010). Bilingual: life and reality.

Guven, C. and Islam, A. (2015). Age at migration, language proficiency, and socioeconomic outcomes: evidence from australia. *Demography*, 52(2):513–542.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. Berkeley, CA: University of California Press.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive psychology*, 21(1):60–99.

Lindemann, K. (2014). The effects of ethnicity, language skills, and spatial segregation on labour market entry success in estonia. *European Sociological Review*, 30(1):35–48.

Lu, S., Chen, S., and Wang, P. (2019). Language barriers and health status of elderly migrants: Micro-evidence from china. *China Economic Review*, 54:94–112.

McManus, W. S. (1985). Labor market costs of language disparity: An interpretation of hispanic earnings differences. *The American Economic Review*, 75(4):818–827.

Meara, P. (1980). Vocabulary acquisition: A neglected aspect of language learning. *Language Teaching*, 13(3-4):221–246.

Miranda, A. and Zhu, Y. (2013). English deficiency and the native-immigrant wage gap. *Economics Letters*, pages 38–41.

Olea, J. L. M. and Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369.

Parker, S. W., Rubalcava, L., and Teruel, G. (2005). Schooling inequality and language barriers. *Economic Development and Cultural Change*, 54(1):71–94.

Patrinos, H. A., Velez, E., and Psacharopoulos, G. (1994). Language, education, and earnings in asuncion, paraguay. *The Journal of developing areas*, pages 57–68.

Shields, M. A. and Price, S. W. (2002). The english language fluency and occupational success of ethnic minority immigrant men living in english metropolitan areas. *Journal of population Economics*, 15:137–160.

Solís, P., Güémez, B., and Campos-Vázquez, R. M. (2025). Skin tone and inequality of socioeconomic outcomes in mexico: A comparative analysis using optical colorimeters and color palettes. *Sociology of Race and Ethnicity*, 11(1):50–68.

StataCorp (2025). Stata 18 base reference manual. Technical report, College Station, TX: Stata Press.

Telles, E., Flores, R. D., and Urrea-Giraldo, F. (2015). Pigmentocracies: Educational inequality, skin color and census ethnoracial identification in eight latin american countries. *Research in Social Stratification and Mobility*, 40:39–58.

Telles, E. and Paschel, T. (2014). Who is black, white, or mixed race? how skin color, status, and nation shape racial classification in latin america. *American Journal of Sociology*, 120(3):864–907.

Wang, H., Smyth, R., and Cheng, Z. (2017). The economic returns to proficiency in english in china. *China Economic Review*, 43:91–104.

Weinreich, U. (2010). *Languages in contact: Findings and problems*. Number 1. Walter de Gruyter.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838.

Yao, Y. and Van Ours, J. C. (2015). Language skills and labor market performance of immigrants in the netherlands. *Labour Economics*, 34:76–85.

Yao, Y. and van Ours, J. C. (2019). Daily dialect-speaking and wages among native dutch speakers. *Empirica*, 46:653–668.

Table 1. Descriptive statistics of the analytical sample. Only individuals who declare being of indigenous ethnicity are included. Mean (SD) reported for continuous variables and number of cases (percentage) for discrete variables.

	Spanish as an additional language (SAL)		
	No (N=637) (49%)	Yes (N=663) (51%)	Test
Female	0.56 (0.50)	0.49 (0.50)	0.011
Age	42.86 (11.26)	43.16 (11.24)	0.626
Father indigenous lang	0.40 (0.49)	0.94 (0.23)	<0.001
Mother indigenous lang	0.40 (0.49)	0.94 (0.24)	<0.001
Years of education	11.54 (4.51)	9.08 (4.89)	<0.001
Qualifications			
Primary	188 (29.5%)	273 (41.2%)	<0.001
Secondary	220 (34.5%)	194 (29.3%)	
Prepa	98 (15.4%)	56 (8.4%)	
Degree	67 (10.5%)	25 (3.8%)	
Other	64 (10.0%)	115 (17.3%)	
Self-assessment PRODER skin tone			
E	134 (21.0%)	143 (21.6%)	0.077
A-B	75 (11.8%)	100 (15.1%)	
C	119 (18.7%)	134 (20.2%)	
D	125 (19.6%)	125 (18.9%)	
F	63 (9.9%)	75 (11.3%)	
G	47 (7.4%)	41 (6.2%)	
H	41 (6.4%)	26 (3.9%)	
I-K	33 (5.2%)	19 (2.9%)	
Current work status	0.69 (0.46)	0.70 (0.46)	0.955
Imputed income	7.74 (3.45)	5.87 (3.14)	<0.001
Number of hh bulbs	6.82 (3.63)	5.52 (3.12)	<0.001
Life satisfaction	8.42 (1.37)	8.16 (1.48)	0.001
Health status			
Regular	206 (32.3%)	229 (34.5%)	0.588
Good	383 (60.1%)	380 (57.3%)	
Very good	48 (7.5%)	54 (8.1%)	
hthserv			
SP	327 (51.3%)	433 (65.3%)	<0.001
IMSS	98 (15.4%)	66 (10.0%)	
Other	212 (33.3%)	164 (24.7%)	

Total sample: N = 1,300.

Table 1. Descriptive statistics of the analytical sample (Cont).

	Spanish as additional language (SAL)		
	No	Yes	Test
Mother qualifications			
None	211 (33.1%)	356 (53.7%)	<0.001
Primary	285 (44.7%)	222 (33.5%)	
Secondary	52 (8.2%)	18 (2.7%)	
Other	89 (14.0%)	67 (10.1%)	
Father qualifications			
None	156 (24.5%)	297 (44.8%)	<0.001
Primary	293 (46.0%)	253 (38.2%)	
Secondary	57 (8.9%)	23 (3.5%)	
Other	131 (20.6%)	90 (13.6%)	
Mother's years of education	8.93 (3.56)	7.17 (2.54)	<0.001
Father's years of education	9.06 (3.85)	7.54 (3.02)	<0.001
Mother occupation			
No work	420 (65.9%)	484 (73.0%)	0.004
Priv employ	27 (4.2%)	10 (1.5%)	
Self-employ	142 (22.3%)	129 (19.5%)	
Other	48 (7.5%)	40 (6.0%)	
fthoccupa			
Priv employ	109 (17.1%)	55 (8.3%)	<0.001
Self-employ	343 (53.8%)	448 (67.6%)	
Other	185 (29.0%)	160 (24.1%)	
Fam oring imputed income (short definition)	1.91 (1.10)	1.46 (1.02)	<0.001

Total sample: N = 1,300.

Table 2. Family Indigenous language background (*inlang*)

	Mother's inlang			Father's inlang		
Own inlang	No	Yes	Tot	No	Yes	Tot
No	385	252	637	383	254	637
	(90.4%)	(28.8%)	(49.0%)	(91.0%)	(28.9%)	(49.0%)
Yes	41	622	663	38	625	663
	(9.6%)	(71.2%)	(51.0%)	(9.0%)	(71.1%)	(51.0%)
Total	426	874	1300	421	879	1300
	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)

Table 3. Mothers vs Fathers indigenous language status (*inlang*)

	Father's inlang		
Mother's inlang	No	Yes	Tot
No	324	102	426
	(77.0%)	(11.6%)	(32.8%)
Yes	97	777	874
	(23.0%)	(88.4%)	(67.2%)
Total	421	879	1300
	(100%)	(100%)	(100%)

Table 4. Regression for wellbeing outcomes on indigenous language (All sample). Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. Montiel Olea-Pflueger (2013) effective first-stage F statistic reported. Robust standard errors in parenthesis. *10% significant; **5% significant; ***1% significant.

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			OLS			
Coeff.	-0.02	-1.11***	-1.18***	-0.82***	-0.16*	0.01
	(0.027)	(0.248)	(0.195)	(0.229)	(0.093)	(0.036)
			2SLS			
Coeff.	0.02	-0.91**	-1.14***	-1.25***	-0.30*	-0.03
	(0.048)	(0.445)	(0.363)	(0.377)	(0.161)	(0.065)
Effective first-stage F	256.04	256.04	256.04	256.04	256.04	256.04
J-stat	0.48	0.18	0.13	2.59	0.01	0.18
Prob $> \chi^2(1)$	0.49	0.67	0.71	0.11	0.91	0.67
Endog. robust score	1.19	0.30	0.02	2.08	1.20	0.64
Prob $> \chi^2(1)$	0.28	0.58	0.88	0.15	0.27	0.43
			MLE			
Coeff.		-0.99**	-1.16***	-1.38***	-0.22	-0.02
		(0.453)	(0.373)	(0.361)	(0.148)	(0.063)
ATE		-0.988**	-1.156***	-1.377***	-0.218	-0.025
		(0.453)	(0.373)	(0.361)	(0.148)	(0.063)
ATET		-0.988**	-1.156***	-1.377***	-0.218	-0.025
		(0.453)	(0.373)	(0.361)	(0.148)	(0.063)
$\rho = 0$ test $\chi^2(1)$		0.11	0.01	4.10	0.33	0.40
Prob $> \chi^2(1)$		0.74	0.93	0.04	0.57	0.53
			CF			
Coeff.		-0.99**	-1.16***	-1.46***	-0.23	-0.02
		(0.436)	(0.355)	(0.391)	(0.159)	(0.062)
ATE		-0.994**	-1.158***	-1.462***	-0.225	-0.023
		(0.436)	(0.355)	(0.391)	(0.159)	(0.062)
ATET		-0.994**	-1.158***	-1.462***	-0.225	-0.023
		(0.436)	(0.355)	(0.391)	(0.159)	(0.062)
$\rho = 0$ test $\chi^2(1)$		0.10	0.01	4.19	0.31	0.38
Prob $> \chi^2(1)$		0.75	0.93	0.04	0.58	0.54

Table 4. (Cont.).

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			PO-MLE			
Coeff.		-1.13**	-1.14***	-3.18***	-0.20	-0.01
		(0.469)	(0.376)	(0.331)	(0.164)	(0.064)
ATE		-1.130**	-1.142***	-3.179***	-0.200	-0.012
		(0.469)	(0.376)	(0.331)	(0.164)	(0.064)
ATET		-0.834*	-0.891**	-1.876***	-0.235	-0.037
		(0.465)	(0.406)	(0.342)	(0.156)	(0.067)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		1.29	1.51	70.83	0.49	0.81
Prob $> \chi^2(2)$		0.52	0.47	0.00	0.78	0.67
			PO-CF under joint normality			
Coeff.		-1.13**	-1.29***	-1.63***	-0.21	-0.01
		(0.467)	(0.379)	(0.429)	(0.162)	(0.063)
ATE		-1.127**	-1.289***	-1.629***	-0.206	-0.011
		(0.467)	(0.379)	(0.429)	(0.162)	(0.063)
ATET		-0.870*	-1.037***	-1.305***	-0.243	-0.035
		(0.445)	(0.361)	(0.387)	(0.165)	(0.065)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		1.18	1.61	4.70	0.48	0.79
Prob $> \chi^2(2)$		0.56	0.45	0.10	0.79	0.67
			PO-CF under linearity of $E(u\|v)$			
ATE	0.01	-1.27**	-1.15***	-1.46***	-0.21	-0.01
	(0.051)	(0.506)	(0.427)	(0.507)	(0.179)	(0.073)
ATET	-0.01	-0.99**	-1.76***	-1.79***	-0.29	-0.06
	(0.052)	(0.507)	(0.400)	(0.409)	(0.173)	(0.073)
Endog. test $\chi^2(1)$	0.73	0.69	0.70	5.39	0.69	1.50
Prob $> \chi^2(1)$	0.69	0.71	0.70	0.07	0.71	0.47
N	1292	1292	1292	1292	1292	1292
N_0	659	659	659	659	659	659
N_1	633	633	633	633	633	633

Table 5. Regression for wellbeing outcomes on indigenous language (Females). Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. Montiel Olea-Pflueger (2013) effective first-stage F statistic reported. Robust standard errors in parenthesis. *10% significant; **5% significant; ***1% significant.

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			OLS			
Coeff.	-0.04	-1.21***	-1.35***	-0.90***	-0.05	0.02
	(0.043)	(0.326)	(0.246)	(0.251)	(0.134)	(0.050)
			2SLS			
Coeff.	0.02	-1.23**	-1.81***	-1.84***	-0.41*	-0.10
	(0.080)	(0.615)	(0.494)	(0.476)	(0.246)	(0.099)
Effective first-stage F	118.00	118.00	118.00	118.00	118.00	118.00
J-stat	0.14	1.28	0.02	0.70	0.03	0.50
Prob $> \chi^2(1)$	0.71	0.26	0.90	0.40	0.87	0.48
Endog. robust score	0.73	0.00	1.13	5.36	3.06	2.16
Prob $> \chi^2(1)$	0.39	0.97	0.29	0.02	0.08	0.14
			MLE			
Coeff.		-1.21*	-1.81***	-1.90***	-0.33	-0.12
		(0.627)	(0.513)	(0.390)	(0.242)	(0.116)
ATE		-1.208*	-1.809***	-1.904***	-0.331	-0.123
		(0.627)	(0.513)	(0.390)	(0.242)	(0.116)
ATET		-1.208*	-1.809***	-1.904***	-0.331	-0.123
		(0.627)	(0.513)	(0.390)	(0.242)	(0.116)
$\rho = 0$ test $\chi^2(1)$		0.00	1.10	11.72	2.29	2.02
Prob $> \chi^2(1)$		0.99	0.29	0.00	0.13	0.15
			CF			
Coeff.		-1.21**	-1.75***	-2.08***	-0.34	-0.09
		(0.593)	(0.464)	(0.480)	(0.247)	(0.091)
ATE		-1.208**	-1.755***	-2.079***	-0.338	-0.091
		(0.593)	(0.464)	(0.480)	(0.247)	(0.092)
ATET		-1.208**	-1.755***	-2.079***	-0.338	-0.091
		(0.593)	(0.464)	(0.480)	(0.247)	(0.092)
$\rho = 0$ test $\chi^2(1)$		0.00	1.07	8.95	2.12	2.08
Prob $> \chi^2(1)$		0.99	0.30	0.00	0.15	0.15

Table 5. (Cont.).

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			PO-MLE			
Coeff.		-1.25*	-1.61***	-2.56***	-0.34	-0.10
		(0.654)	(0.462)	(0.746)	(0.245)	(0.102)
ATE		-1.249*	-1.609***	-2.561***	-0.339	-0.095
		(0.654)	(0.462)	(0.746)	(0.245)	(0.102)
ATET		-1.178*	-1.877***	-2.071***	-0.316	-0.179
		(0.634)	(0.513)	(0.474)	(0.261)	(0.123)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		0.04	1.92	10.14	2.65	3.38
Prob $> \chi^2(2)$		0.98	0.38	0.01	0.27	0.18
			PO-CF under joint normality			
Coeff.		-1.24**	-1.67***	-2.13***	-0.48**	-0.06
		(0.612)	(0.483)	(0.509)	(0.244)	(0.093)
ATE		-1.235**	-1.667***	-2.127***	-0.485**	-0.063
		(0.613)	(0.483)	(0.510)	(0.244)	(0.093)
ATET		-1.176*	-1.860***	-2.073***	-0.477*	-0.125
		(0.614)	(0.477)	(0.489)	(0.255)	(0.098)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		0.04	1.92	9.44	5.21	3.37
Prob $> \chi^2(2)$		0.98	0.38	0.01	0.07	0.19
			PO-CF under linearity of $E(u\|v)$			
ATE	-0.04	-1.49*	-1.36**	-2.27***	-0.38	-0.08
	(0.096)	(0.766)	(0.606)	(0.649)	(0.276)	(0.118)
ATET	-0.02	-0.96	-2.31***	-2.22***	-0.45	-0.13
	(0.085)	(0.697)	(0.549)	(0.547)	(0.277)	(0.109)
Endog. test $\chi^2(1)$	0.22	0.45	3.62	9.60	1.94	3.38
Prob $> \chi^2(1)$	0.90	0.80	0.16	0.01	0.38	0.18
N	681	681	681	681	681	681
N_0	323	323	323	323	323	323
N_1	358	358	358	358	358	358

Table 6. Regression for wellbeing outcomes on indigenous language (Males). Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. Montiel Olea-Pflueger (2013) effective first-stage F statistic reported. Robust standard errors in parenthesis. *10% significant; **5% significant; ***1% significant.

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			OLS			
Coeff.	0.00	-1.11***	-1.08***	-0.70*	-0.29**	0.00
	(0.032)	(0.376)	(0.301)	(0.378)	(0.129)	(0.054)
			2SLS			
Coeff.	0.01	-0.73	-0.66	-0.70	-0.18	0.07
	(0.054)	(0.603)	(0.504)	(0.547)	(0.201)	(0.085)
Effective first-stage F	141.27	141.27	141.27	141.27	141.27	141.27
J-stat	0.64	0.04	0.04	0.81	0.11	0.27
Prob $> \chi^2(1)$	0.42	0.85	0.85	0.37	0.74	0.61
Endog. robust score	0.11	0.56	1.30	0.00	0.54	1.09
Prob $> \chi^2(1)$	0.75	0.45	0.25	1.00	0.46	0.30
			MLE			
Coeff.		-1.00	-0.73	-0.87	-0.14	0.08
		(0.612)	(0.518)	(0.541)	(0.205)	(0.079)
ATE		-1.004	-0.725	-0.865	-0.138	0.075
		(0.612)	(0.518)	(0.541)	(0.205)	(0.079)
ATET		-1.004	-0.725	-0.865	-0.138	0.075
		(0.612)	(0.518)	(0.541)	(0.205)	(0.079)
$\rho = 0$ test $\chi^2(1)$		0.05	0.89	0.24	0.94	1.58
Prob $> \chi^2(1)$		0.83	0.34	0.63	0.33	0.21
			CF			
Coeff.		-1.13*	-0.75	-0.86	-0.13	0.08
		(0.593)	(0.497)	(0.539)	(0.201)	(0.082)
ATE		-1.132*	-0.755	-0.860	-0.128	0.080
		(0.593)	(0.497)	(0.539)	(0.201)	(0.082)
ATET		-1.132*	-0.755	-0.860	-0.128	0.080
		(0.593)	(0.497)	(0.539)	(0.201)	(0.082)
$\rho = 0$ test $\chi^2(1)$		0.00	0.83	0.18	1.08	1.45
Prob $> \chi^2(1)$		0.96	0.36	0.67	0.30	0.23

Table 6. (Cont.).

SAL	work	yrsedu	iincome	bulbs	lifesat	hthstat
			PO-MLE			
Coeff.		-1.21*	-0.89	-3.10***	0.57	0.07
		(0.671)	(0.597)	(0.561)	(0.557)	(0.085)
ATE		-1.207*	-0.889	-3.101***	0.571	0.072
		(0.671)	(0.597)	(0.561)	(0.557)	(0.085)
ATET		-0.865	-0.308	-1.277**	0.004	0.074
		(0.634)	(0.592)	(0.649)	(0.310)	(0.081)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		0.57	3.69	21.38	1.24	1.53
Prob $> \chi^2(2)$		0.75	0.16	0.00	0.54	0.47
			PO-CF under joint normality			
Coeff.		-1.18*	-1.07**	-1.16*	-0.05	0.08
		(0.691)	(0.528)	(0.645)	(0.212)	(0.087)
ATE		-1.182*	-1.069**	-1.155*	-0.048	0.078
		(0.692)	(0.529)	(0.645)	(0.212)	(0.087)
ATET		-0.914	-0.581	-0.678	-0.173	0.081
		(0.604)	(0.505)	(0.546)	(0.205)	(0.085)
$\rho_0 = \rho_1 = 0$ test $\chi^2(2)$		0.48	4.43	1.58	2.32	1.45
Prob $> \chi^2(2)$		0.79	0.11	0.45	0.31	0.48
			PO-CF under linearity of $E(u\|v)$			
ATE	0.00	-1.49**	-1.17**	-0.89	0.00	0.06
	(0.040)	(0.635)	(0.541)	(0.647)	(0.227)	(0.089)
ATET	-0.02	-1.21*	-1.53***	-1.20**	-0.20	0.01
	(0.039)	(0.701)	(0.546)	(0.576)	(0.208)	(0.095)
Endog. test $\chi^2(1)$	0.69	0.40	0.26	0.03	2.48	1.00
Prob $> \chi^2(1)$	0.71	0.82	0.88	0.99	0.29	0.61
N	611	611	611	611	611	611
N_0	336	336	336	336	336	336
N_1	275	275	275	275	275	275

Online appendix for "The causal effect of speaking Spanish as an additional language on education, labor, and wellbeing outcomes, among the indigenous ethno-linguistic minorities of Mexico"

Alfonso Miranda [a,b,1,]

[a]*Applied Economics Division, CIDE, Mexico.*
[b]*Institute for the Study of Labor (IZA), Germany*

Abstract

Here we offer further technical detail of the data and methods used in the paper. Further regression results and robustness checks are also reported.

1. Data

1.1. PRODER 2019 *sample design and covered topics*

PRODER uses INEGI's geostatistical frame for sampling; which is a list of basic geostatisitical areas (AGEBs) each with information on population and detailed cartography.[1] In the urban domain AGEBs are divided in localities and localities in blocks (*manzanas*). In the rural domain AGEBs are divided simply in localities. The target population were 25 to 64 Mexican individuals living in households between July 30th and October 11th 2019 in Mexico. Computer assisted personal interviewing (CAPI) was implemented—mainly using smart phones. Interviews were all done in Spanish.

To achieve national representativity as well as securing enough variation for the ethnoracial characteristics of interest, PROCER's sample is composed by a series of sub-samples. There is, first, a national representative sub-sample of 3,187 individuals. To this, five regional representative sub-samples, of 800 individuals each, are added for: (i) Metropolitan area of Mexico City; (ii) Metropolitan area of Monterrey; (iii) Metropolitan area of Oaxaca; (iv) Metropolitan area of Merida; and (v) Maya area.[2] It was noted, however, that part of the 800 sample in each region

[1]INEGI is the Spanish acronym for the National Institute for Statistics and Geography; equivalent to the Census Bureau in the USA, Statistics Canada, or the National Office for Statistics in the UK.

[2]This region includes the muncipalities of: Hopelchén, Calakmul, José María Morelos, Cantamayec, Chacsinkín,

could be *"donated"* (transferred) to the national sample in the understanding that by chance some observations will fall in the region. After allocation, it was determined that a total of 883 observations from the regional sub-samples could be donated to the national sample: 666 from Mexico City, 147 from Monterrey, 20 from Oaxaca, 40 from Merida, and 10 from the Maya Zone. Taken together, national representative sub-sample plus donations, give 4070 units. This is called the 'national' sample. Next, within each region, the 'complement' sub-sample is obtained subtracting donations from 800. In Monterrey, for instance, the complement sub-sample has 653 units. Summing up all complement sub-samples $3,117$ units are obtained. This is called the 'complement' sample. Finally, 'national + complement' gives PROCER's total sample of $N = 7,187$.

The sample has a stratified multistage cluster design. There are six strata: Five for the (i)-(v) regions and one for the national sub-sample. Proportional allocation is used at each selection stage. AGEBs are the primary sampling units (PSUs); selected systematically with probability proportional to size (PPS) within each stratum. Blocks (urban domain) and localities (rural domain) are the secondary sampling units (SSUs), also selected systematically with probability proportional to size. Finally, households and individuals are selected with equal probability in the third stage.

The questionnaire is composed by eight sections: (a) Ethnoracial characteristics; (b) basic socio-demographics; (c) extended socio-demographics at the time of the interview and when the subject was 14; (d) Perceived discrimination; (e) Opinion about ethnoracial inequality and access to public services; (f) Social capital; (g) Ethnoracial and socio-demographic characteristics of the partner; (h) Health status and access to public health services.

1.2. Permanent income of the origin family

While PRODER does not does collect information about parents' income, there is however a module about subjects' 'socio-economic origins' where PRODER inquires about different aspects of their lives when aged 14. In this module there are a series of dichotomous questions about the presence of services and appliances at the origin household: (1) pipe water inside the dwelling, (2) electric or gas stove, and (3) electricity. If there was electricity then there are follow-up questions about the presence of: (4) tv, (5) fridge, (6) washing machine, (7) blender, (8) radio, tape recorder, and/or CD player, (9) electric toaster (10) land line. Adding up these responses we build an imputed permanent income of the family of origin's that ranges from 0 to 10. Unfortunately, upon tabulation (see table 1), we find that this variable is missing for 26% of the analytical sam-

Chankom, Chikindzonot, Maní, Mayapán, Ozkutzcab, Tahdziú, Teabo, Tekom, Tixcacalcupul, Tixmehuac y Yaxcabá.

ple; which make us think it suffers from substantial non-recall. In practice the analytical sample size goes from $N = 1,300$ to only 955. Given that $N = 1,300$ is not a particularly large sample, we find that using this variable as control would imply an unacceptable loss of variation. More importantly, missingness status is quite different in the $SAL = 1$ (32%) sample than in the $SAL = 0$ sample (21%). Hence, it is unlikely missingness is at random. We conclude, therefore, that this is not a good control for analysis.

We propose instead using questions (1)-(3) to build our imputed permanent income of the origin family. We call this variable the 'short definition' of the imputed permanent income of the family of origin. Now only 8 observations have missing value and there is no evidence that missingness status is different among SAL and non-SAL individuals. We believe this is a good control for analysis.

1.3. Access to health services

Mexico does not have a universal national health system but rather various subsystems that individuals access depending on whether they work for the (formal) private sector, for the public sector, for PEMEX (state's oil company), are self-eployed in the informal sector, or do not work at all. Those who work in the private sector have access to IMSS, those who work for the public sector have access to ISSSTE (federal or state service) and those who work for PEMEX have access to their own health institute. Self-employed individuals have access to Seguro Popular (SP). Finally, there is the private health sector and other special subsystems such as the health institute for the army.

Regarding our analytical sample, $SAL = 1$ individuals have a lower probability (10%) to have IMSS access than $SAL = 0$ (15%) individuals. In contrast, individuals in the $SAL = 1$ sample have a much higher probability to have SP access (65%) than individuals in the $SAL = 0$ sample (51%). These stylized facts conform with our prior expectations.

2. Econometric methods

2.0.1. Endogenous treatment probit and Endogenous treatment ordinal probit models

Up to now we estimated models for continuous response variables for all the variables under investigation. Current work status is, however, a binary response while health status is an ordinal response. For completenes we estimate endogenous treatment probit (ETP) and endogenous ordinal probit (ETOP) models for these two responses.

The structural model for the ETP is

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i\beta + \theta t_i + a_s + u_i, && (1)\\
t_i^* &= \mathbf{w}_i\gamma + b_s + v_i, && (2)\\
y_i &= \mathbb{1}\left(y_i^* > 0\right) && (3)\\
t_i &= \mathbb{1}\left(t_i^* > 0\right) && (4)
\end{aligned}
$$

with $(u,v) \sim MVN(\mathbf{0},\Sigma)$. The model is fitted by maximum likelihood. Define $q_y = 2y-1$, $q_t = 2t-1$, $\varepsilon_y = \mathbf{x}_i\beta + \theta t_i + a_s$, and $\varepsilon_t = \mathbf{w}_i\gamma + b_s$. The contribution of the i-th individual to the log-likelihood is given by

$$\log L = \sum_i \log \Phi_2\left(q_{yi}\varepsilon_{yi}, q_{ti}\varepsilon_{ti}, q_{yi}q_{ti}\rho\right),$$

where $\Phi_2(\cdot)$ represents the bivariate normal cumulative distribution function and ρ is the correlation coefficient between y_i and t_i. Given that both responses are binary, the variances are only identified up to a constant; both set to one. Upon maximization $-\widehat{H}^{-1}$ provides a consistent estimator of the covariance matrix. A robust estimator (Huber-White) of the covariance matrix is available forming the usual sandwich estimator.

To accommodate an ordinal response that takes on values $1,..,H$ equation (2) is modified to specify a model with $H+1$ thresholds,

$$
y_i = \begin{cases}
1, & \text{if } \kappa_0 < y^* \leq \kappa_1,\\
2, & \text{if } \kappa_1 < y^* \leq \kappa_2,\\
\vdots & \vdots\\
H, & \text{if } \kappa_{H-1} < y^* \leq \kappa_H.
\end{cases}
$$

with $\kappa_0 = -\infty$ and $\kappa_H = \infty$. To write the log-likelihood we will use a bivariate normal distribution and carefully define the lower and upper limits of integration. Define,

$$
l_{yi} = \begin{cases}
-\infty, & \text{if } y_i = 0,\\
\kappa_{y_i-1} - \mathbf{x}_i\beta - \theta t_i, & \text{if } y_i = 1,\ldots,H-1,\\
\infty, & \text{if } y = H.
\end{cases}
\quad ; \; u_{yi} = \begin{cases}
-\infty, & \text{if } y_i = 0,\\
\kappa_{y_i} - \mathbf{x}_i\beta - \theta t_i, & \text{if } y_i = 1,\ldots,H-1,\\
\infty, & \text{if } y = H.
\end{cases}
$$

and

$$l_{ti} = \begin{cases} -\infty, & \text{if } t_i = 0, \\ -\mathbf{w}_i\gamma, & \text{if } t_i = 1 \end{cases} ; \; u_{yi} = \begin{cases} -\mathbf{w}_i\gamma, & \text{if } t_i = 0, \\ \infty, & \text{if } t_i = 1. \end{cases}$$

then the log-likelihood can be written as

$$\log L = \sum_i \log \left[\int_{l_{yi}}^{u_{yi}} \int_{l_{ti}}^{u_{ti}} \phi_2(\varepsilon, \Sigma) \, d\varepsilon_y d\varepsilon_t \right],$$

where $\phi_2(\varepsilon, \Sigma)$ is the density of a bivariate normal $\varepsilon = (\varepsilon_y, \varepsilon_t)'$ with mean vector zero and variance

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

As usual, the model is fitted by maximum likelihood. At convergence $-\widehat{H}^{-1}$ provides a consistent estimator of the covariance matrix. A robust estimator (Huber-White) of the covariance matrix is available forming the usual sandwich estimator.

2.0.2. Instrumental variables quantile regression

To finalize we estimate an smoothed estimating equations instrumental variables regression (SEE-IVQR) as suggested by Kaplan and Sun (2017) and implemented by command `ivqregress` in Stata. Here we provide only a brief, intuitive, description of the SEE-IVQR estimator avoiding as much technical detail as possible. The SEE-IVQR estimator considers a random coefficients model for a continuous response y

$$y_i = \mathbf{x}_i\beta(u) + \theta(u)t_i$$

where $\mathbf{x}$ is a vector of exogenous explanatory variables, t is the endogenous treatment indicator, and u is a random variable that characterizes the heterogeneity and all unobservables that affect y. $\beta(u)$ and $\theta(u)$ are conformable vectors of random coefficients. Notice that, because t is endogenous, $\theta(u)$ is generally a function of the exogenous variables $\mathbf{x}$, a set of instrumental variables $\mathbf{z}$, and an error term v that is correlated with u. Under some regularity conditions the model implies the conditional probability

$$P[y_i \leq \mathbf{x}_i\beta(\tau) + \theta(\tau)t_i | \mathbf{z}, \mathbf{x}] = \tau$$

where τ represents a given quantile of y. This conditional probability, in turn, implies the conditional moment condition,

$$E\left[\tau - \mathbb{1}\left(y_i \leq \mathbf{x}_i\beta(\tau) + \theta(\tau)t_i\right)|\mathbf{x},\widehat{t}\right] = 0, \tag{5}$$

where $\widehat{t}$ is the linear projection of t on $\mathbf{x}$ and $\mathbf{z}$—which can be used as instrument for t. Condition (5) defines a set of nonlinear estimating equations that, abstracting for the fact that the indicator function $\mathbb{1}(\cdot)$ is nonconvex and nonsmooth, can be solved to obtain estimators $\widehat{\beta}(\tau)$ and $\widehat{\theta}(\tau)$. This cannot be done easily, however, due to the optimization problems that arise from the properties of $\mathbb{1}(\cdot)$. To get around, Kaplan and Sun (2017) suggest smoothing condition (5) using a kernel function $G(\cdot)$,

$$E\left[\tau - G\left(y_i \leq \mathbf{x}_i\beta(\tau) + \theta(\tau)t_i\right)|\mathbf{x},\widehat{t}\right] = 0,$$

which now define a set of somooth nonlinear equations that can be solved using standard optimization tools. Notice that at each quantile τ a local average treatment effect (LATE) is delivered from this estimator.

3. Further results

3.1. Non-linear models

Table 4 presents results from an endogenous treatment logit regression for current work status—a binary response—for the whole sample. Similarly, table 5 presents results from an endogenous treatment ordinal probit for health status—an ordinal response—for the whole sample.

For current work status (*work*) the coefficient on the *SAL* treatment is statistically insignificant no matter if treatment endogeneity is accounted for or not. This is consistent with the 2SLS results reported in the main text. A test for the exclusion of the instruments on treatment equation rejects the null with a $\chi^2 = 317.2$ ($p - val < 0.001$). In other works, the instruments are strong predictors of treatment status. So, the endogenous treatment probit model is well identified. Moreover, a test for $\rho = 0$ fails to reject the null at 1%; which implies that the treatment is in fact exogenous. In summary, we are confident that that the *SAL* treatment is exogenous and has null effect on current work status.

Moving to the endogenous treatment ordinal probit for health status (*hthstat*) in table 5 we find similar results: *SAL* is exogenous and has null effect on health status.

3.2. Instrumental variables quantile regression

Table 6 presents results from instrumental variables quantile regression for years of education, imputed permanent income, bulbs, and life satisfaction—all the continuous variables considered in the present study. Quantiles Q_{25}, Q_{50}, and Q_{75} are considered. Coefficients in these models are interpreted as a LATE at the corresponding conditional quantile. Here we only refer to results for the whole sample.

For years of education we find no significant effect at quantiles Q_{25} and Q_{50}. However, the effect is significant at the top of the distribution at Q_{75}, where the LATE is found to be -1.27 and significant at 1%. This falls within the 95% confidence interval of the OLS point estimate. Hence, evidence suggest that *SAL* status affects years of education particularly at the top of the education distribution. However, diagnostics reported in table 7 fail to reject the null of a constant effect; which suggests that the researcher should keep only a model for the conditional mean. Here, as well, a test for the exogeneity of the treatment fails to reject the null.

For imputed permanent income we do detect a significant LATE at all quantiles considered. While the point estimate is -1.06 at Q_{25} (significant at 5%), the point estimate at Q_{75} is -1.27 (significant at 1%). Hence, evidence suggests that the effect *SAL* status is stronger at the top of the income distribution. However, again, diagnostics in table 7 fail to reject the null of a constant effect. Similarly, a test for the exogeneity of the treatment fails to reject the null. Similar conclusions can be drawn for the number of bulbs in the household, which is also a proxy for income.

Finally, for life satisfaction, we find point estimates for the LATE that are statistically insignificant at all considered quantiles. In fact, our QR point estimates fall with the 95% confidence interval of the OLS point estimate. Again, diagnostic tests fail to reject the null of a constant effect as well as the null of exogenous treatment.

References

Kaplan, D. M. and Sun, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, 33(1):105–157.

Table 1. Family of origin imputed permanent income (*iincome*)

	Spanish as additional language		
Missing family of origin imputed income	No	Yes	Tot
No	504	451	955
	(79.1%)	(68.0%)	(73.5%)
Yes	133	212	345
	(20.9%)	(32.0%)	(26.5%)
Total	637	663	1300
	(100%)	(100%)	(100%)

Table 2. Family of origin imputed permanent income (*iincome*) (short definition)

	Spanish as additional language		
Missing family of origin imputed income short	No	Yes	Tot
No	633	659	1292
	(99.4%)	(99.4%)	(99.4%)
Yes	4	4	8
	(0.6%)	(0.6%)	(0.6%)
Total	637	663	1300
	(100%)	(100%)	(100%)

Table 3. First stage of 2SLS regressions. Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. Montiel Olea-Pflueger (2013) effective first-stage F statistic reported. Robust standard errors in parenthesis. *10% significant; **5% significant; ***1% significant.

SAL	Work	Yrsedu	iIncome	Bulbs	Lifesat	Hthstat
			All sample			
Mother indigenous lang	0.34***	0.34***	0.34***	0.34***	0.34***	0.34***
	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)	(0.034)
Father indigenous lang	0.33***	0.33***	0.33***	0.33***	0.33***	0.33***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Effective first-stage F	256.04	256.04	256.04	256.04	256.04	256.04
R^2	0.46	0.46	0.46	0.46	0.46	0.46
N	1292	1292	1292	1292	1292	1292
			Females			
Mother indigenous lang	0.35***	0.35***	0.35***	0.35***	0.35***	0.35***
	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
Father indigenous lang	0.27***	0.27***	0.27***	0.27***	0.27***	0.27***
	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)
Effective first-stage F	118.00	118.00	118.00	118.00	118.00	118.00
R^2	0.46	0.46	0.46	0.46	0.46	0.46
N	681	681	681	681	681	681
			Males			
Mother indigenous lang	0.32***	0.32***	0.32***	0.32***	0.32***	0.32***
	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)	(0.058)
Father indigenous lang	0.42***	0.42***	0.42***	0.42***	0.42***	0.42***
	(0.057)	(0.057)	(0.057)	(0.057)	(0.057)	(0.057)
Effective first-stage F	141.27	141.27	141.27	141.27	141.27	141.27
R^2	0.50	0.50	0.50	0.50	0.50	0.50
N	611	611	611	611	611	611

Table 4. Probit regressions for Work status. Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. ETP stands for endogenous treatment probit. *10% significant; **5% significant; ***1% significant.

SAL	All sample	Females	Males
Probit	-0.07	-0.11	0.02
	(0.090)	(0.114)	(0.150)
ETP	0.06	0.09	0.05
	(0.152)	(0.210)	(0.230)
Instruments ex. test $\chi^2(2)$	317.24	138.30	210.07
Prob $> \chi^2(2)$	<0.001	<0.001	<0.001
$\rho = 0$ test $\chi^2(1)$	1.12	1.44	0.03
Prob $> \chi^2(1)$	0.291	0.230	0.869

Table 5. Ordinal Probit regressions for Health Status. Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects. OP stands for ordered probit, while $ETOP$ standas for endogenous treatment ordered probit. *10% significant; **5% significant; ***1% significant.

SAL	All sample	Females	Males
OP	0.02	0.04	0.01
	(0.076)	(0.105)	(0.112)
ETOP	-0.05	-0.25	0.17
	(0.133)	(0.238)	(0.170)
Instruments ex. test $\chi^2(2)$	317.75	141.72	211.04
Prob $> \chi^2(2)$	<0.001	<0.001	<0.001
$\rho = 0$ test $\chi^2(1)$	0.39	2.16	1.64
Prob $> \chi^2(1)$	0.533	0.141	0.201

Table 6. Quantile regressions of selected responses on Spanish as additional language. Coefficient on SAL each row. Label in the row indicates the response variable. Only individuals who declare being of indigenous ethnicity are included in the analytical sample. Controls include: sex, age, proder skin tone scale, mother's education, father's education, mother's occupation, father's occupation, origin family imputed income, and state fixed-effects.. Robust standard errors in parenthesis. *10% significant; **5% significant; ***1% significant.

Response	Q_{25}	Q_{50}	Q_{75}
		All sample	
yrsedu	-0.61	-0.80	-1.27***
	(0.603)	(0.513)	(0.476)
iincome	-1.06**	-1.35***	-1.47***
	(0.470)	(0.405)	(0.510)
bulbs	-0.78**	-1.23***	-1.72***
	(0.310)	(0.391)	(0.477)
lifesat	-0.17	-0.22	-0.23
	(0.169)	(0.175)	(0.232)
		Females	
yrsedu	-0.63	-1.28**	-1.74**
	(0.848)	(0.562)	(0.822)
iincome	-1.15*	-1.43**	-1.91***
	(0.588)	(0.579)	(0.685)
bulbs	-1.02**	-1.15**	-1.83***
	(0.429)	(0.508)	(0.591)
lifesat	-0.51	-0.41	-0.14
	(0.461)	(0.266)	(0.354)
		Males	
yrsedu	-0.85	-0.45	-1.40*
	(0.898)	(0.695)	(0.781)
iincome	-0.65	-1.31**	-0.91
	(0.814)	(0.636)	(0.892)
bulbs	-0.59	-1.17**	-1.85***
	(0.492)	(0.564)	(0.717)
lifesat	-0.07	-0.11	-0.30
	(0.232)	(0.217)	(0.294)

Table 7. Quantile regressions diagnostics. Test for exogeneity (exog) (critical value (exog_cv)) as well as test for constant effect (conseff) (critical value (conseff_cv)) reported.

	exog	exog_cv	conseff	conseff_cv	N
			All sample		
yrsedu	1.25	2.30	1.04	2.20	1292
iincome	0.74	2.29	0.72	1.98	1292
bulbs	2.14	2.56	2.00	2.37	1292
lifesat	1.36	2.20	0.35	2.03	1292
			Females		
yrsedu	0.76	2.02	1.14	2.16	681
iincome	0.66	2.45	1.04	1.75	681
bulbs	2.30	2.22	1.37	1.93	681
lifesat	1.42	2.65	0.79	1.85	681
			Males		
yrsedu	0.76	2.29	0.56	2.24	611
iincome	0.88	2.15	1.00	1.83	611
bulbs	0.79	2.31	1.80	2.52	611
lifesat	0.68	2.51	0.75	2.18	611

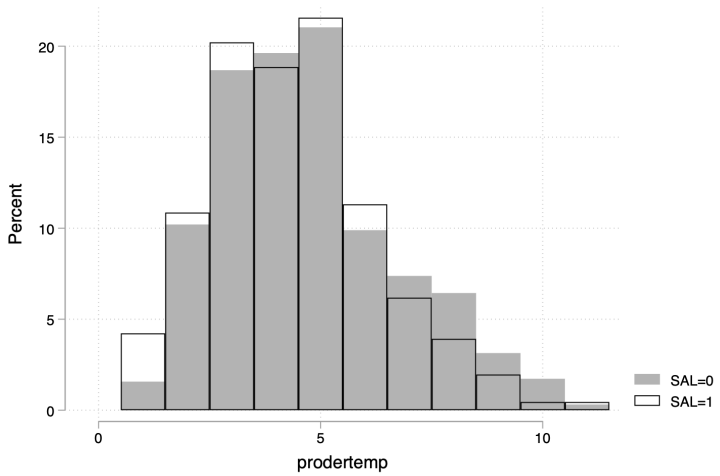

Figure 1. Distribution PRODER skin tone

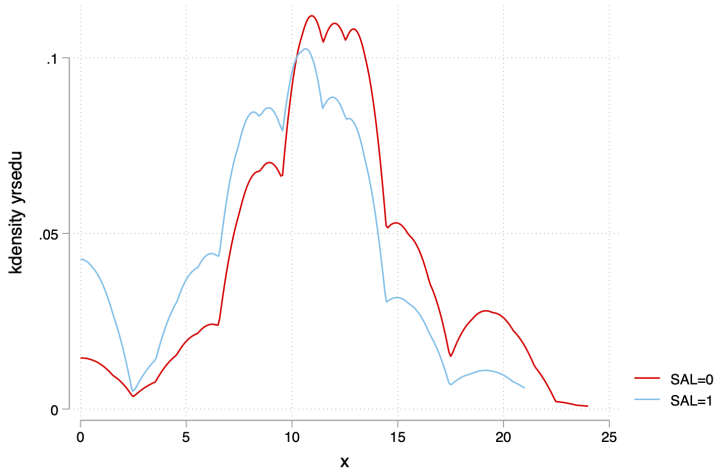

Figure 2. Distribution of years of education

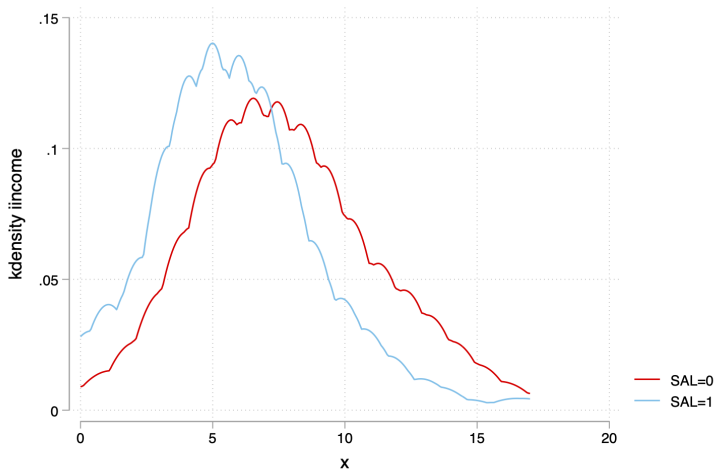

Figure 3. Distribution of imputed income

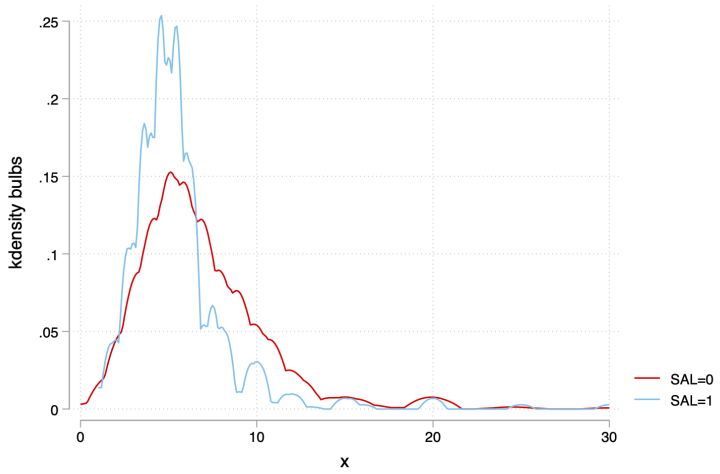

Figure 4. Distribution of number of bulbs

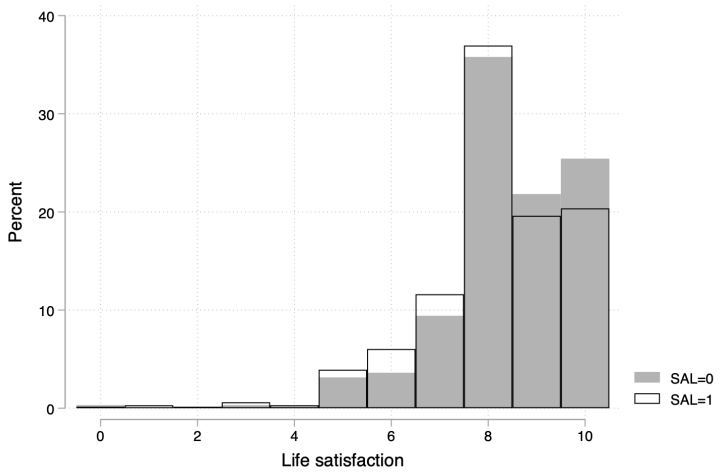

Figure 5. Distribution of life satisfaction

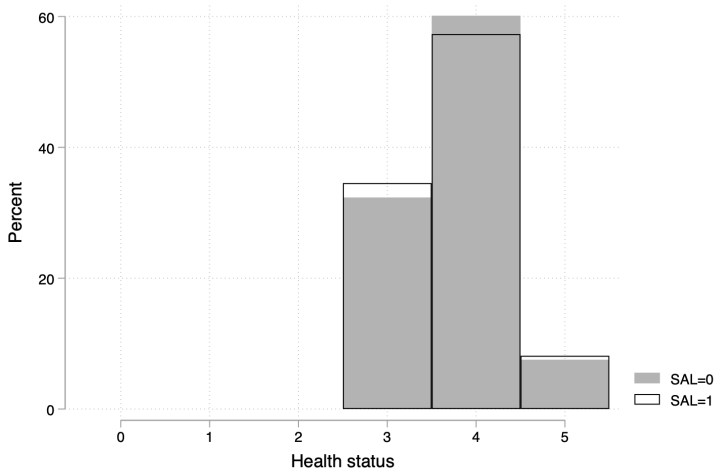

Figure 6. Distribution of health status

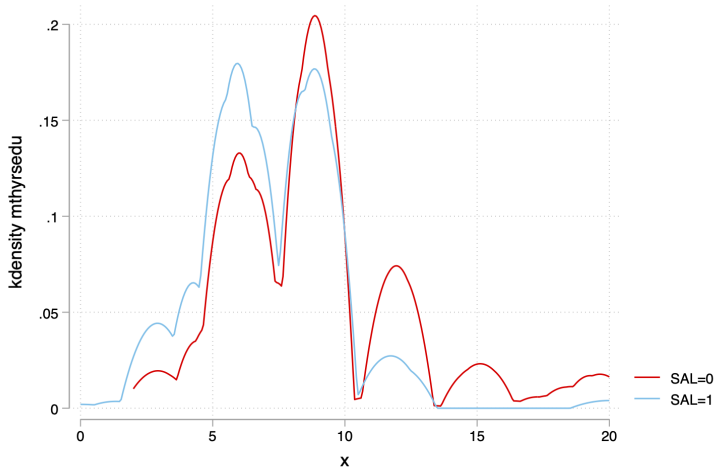

Figure 7. Distribution of mother's years of education

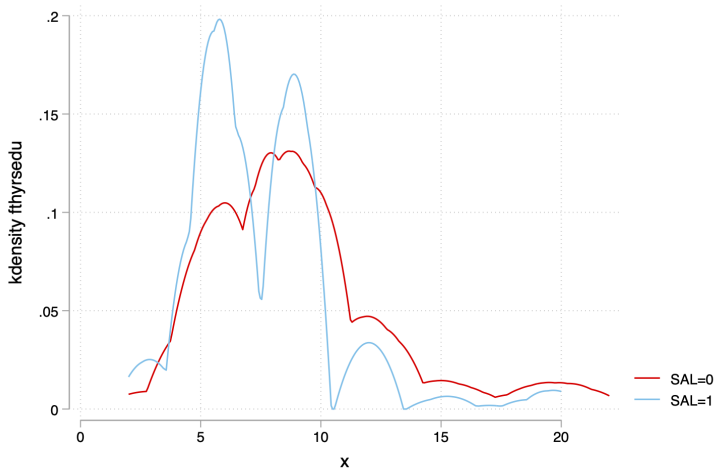

Figure 8. Distribution of father's years of education

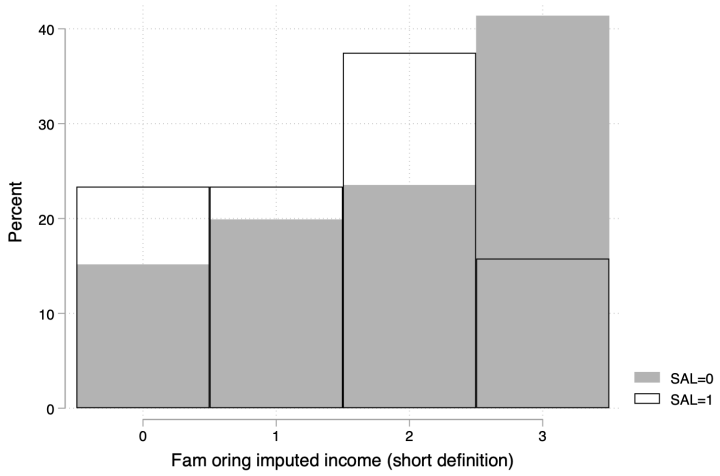

Figure 9. Distribution of origin family income (short definition)