# Online Appendix

## The Marginal Returns to Distance Education: Evidence from Mexico's Telesecundarias

**Emilio Borghesan and Gabrielle Vasey**

### A. Analysis of Cheating and Differential Test-Taking

In this section, we show that the pattern of test scores shown in Figure 1 of the main text is unlikely to be affected by either differential rates of cheating or differential rates of test-taking across school types.

The ENLACE exams are low-stakes, but the Secretariat of Education (SEP) nevertheless attempted to reduce cheating in a number of ways. External administrators oversaw the administration of the exam at each school. Teachers were forbidden from supervising students taking the exams, and exams were graded at a central location. Whenever possible, test-takers were arranged so that they sat adjacent to students in different grades who were taking different exams (de Hoyos, Estrada, and Vargas 2021).

In addition, the SEP developed a statistical tool to determine whether students were likely to have cheated on the exam. This method, the k-index method, identifies students in the same testing location with the same number of correct answers. For all these students, the k-index is calculated as the probability that a group of students have $n_1$ identical incorrect answers given that they have $n_2$ total incorrect answers. Students with a k-index value below a particular threshold are marked as likely having copied.

We find that cheating is slightly more prevalent at telesecundarias than traditional schools, although rates of copying are low at both schools: 1.41% (1.11%) of students at telesecundarias (traditional schools) are deemed to have cheated on their 7th grade exams. The analysis in this paper has excluded those students whose records indicate that they likely cheated. Table A-1 compares the mean scores in each grade in the estimation sample for this paper ("Omitting Copiers") with the larger sample that includes students who likely cheated. The increase in test scores over time in telesecundaria schools relative to traditional schools is

Table A-1
Mean Scores on Exams, by School Type

| | Math | | | | Spanish | | | |
|---|---|---|---|---|---|---|---|---|
| | All | Traditional | Telesecundaria | Difference | All | Traditional | Telesecundaria | Difference |
| Including Copiers | | | | | | | | |
| Grade 6 | 521.4 | 531 | 484.2 | -46.7 | 516.3 | 526.7 | 476.1 | -50.5 |
| Grade 7 | 500.1 | 501.5 | 494.8 | -6.7 | 499.7 | 503.6 | 484.7 | -18.9 |
| Grade 8 | 508.5 | 505 | 522.1 | 17.1 | 477.2 | 479.2 | 469.4 | -9.8 |
| Grade 9 | 524.5 | 518.7 | 546.9 | 28.2 | 499.6 | 501.2 | 493.6 | -7.6 |
| Omitting Copiers | | | | | | | | |
| Grade 6 | 521.1 | 530.7 | 484.0 | -46.7 | 516.1 | 526.5 | 476 | -50.4 |
| Grade 7 | 499.7 | 501.2 | 494.0 | -7.2 | 499.5 | 503.4 | 484.2 | -19.2 |
| Grade 8 | 508.3 | 504.9 | 521.6 | 16.7 | 477 | 479 | 469.1 | -9.9 |
| Grade 9 | 524.3 | 518.5 | 546.6 | 28.1 | 499.5 | 501.1 | 493.3 | -7.7 |

The table shows mean scores on the math and Spanish ENLACE exams for public school students in Mexico for two samples. The sample that omits copiers is the main estimation sample that has been used throughout the paper. The first panel adds students who likely cheated on their exams to this sample. Grade 6 is primary school. Starting in grade 7, public school students attend either traditional secondary schools or telesecundarias. The standard deviations on the math ENLACE exams are 118.3, 98.1, 101.2, and 116.4 in grades 6, 7, 8 and 9, respectively, for this sample. The standard deviations on the Spanish ENLACE exams are 103.9, 97.5, 105.8, and 98.1 in grades 6, 7, 8 and 9, respectively.

minimally affected by the inclusion of copiers.

We next consider whether there are differential rates of test-taking across school types. A pattern whereby students at telesecundarias improve their scores relative to students at traditional schools could occur even in the absence of increased learning at telesecundarias if low-performing students at these schools were less likely to sit the exams. In Table A-2, we investigate this possibility by running several probit regressions. The dependent variable is a binary variable that takes the value of one if a student who is enrolled in the seventh grade sits the ENLACE exam and zero otherwise. In column (1), we include only a dummy for telesecundaria attendance. Column (2) adds controls for students' standardized sixth grade math and Spanish scores. Column (3) includes interactions between telescundaria attendance and sixth grade test scores, and column (4) includes additional controls for the student's age, number of siblings, sex, prospera status, family income, mother's education, number of books in the home, and whether the family has access to a computer.

The results in column (1) show that students who attend telesecundarias are no more or less likely to sit the seventh grade ENLACE exam. Adding covariates for prior exam scores in math and Spanish reveals that telesecundaria students are actually more likely to sit the exams than are students at traditional schools, conditional on past academic performance. Column (3) shows that, for students who attend telesecundarias, the effect of scoring an additional standard deviation higher on their sixth grade math exam is to increase the probability of sitting the seventh grade exam by 0.8 percentage points. Scoring one standard deviation higher on the Spanish sixth grade exam has an average effect of reducing the probability of sitting the exam by 1.4 percentage points. These effects are small given the large change in scores considered, and they are opposite in sign. As a result, it seems unlikely that differential exam-sitting at telesecundarias relative to traditional schools can explain the large gains in test scores for telescundaria students documented in Figure 1 in the main text.

Table A-2
Probability of Sitting Seventh Grade ENLACE Exam

| | Average Derivative | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Tele | 0.0036 | 0.013 | 0.010 | 0.025 |
| | (0.002) | (0.002) | (0.002) | (0.002) |
| Math Score (6th Grade) | | 0.000 | -0.001 | -0.002 |
| | | (0.001) | (0.001) | (0.001) |
| Spanish Score (6th Grade) | | 0.021 | 0.024 | 0.019 |
| | | (0.001) | (0.001) | (0.001) |
| Tele × Math | | | 0.008 | 0.007 |
| | | | (0.003) | (0.003) |
| Tele × Spanish | | | -0.014 | -0.016 |
| | | | (0.003) | (0.003) |
| Controls | No | No | No | Yes |
| Observations | 145908 | 145908 | 145908 | 145908 |

The table shows the average marginal effect of telesecundaria attendance on the probability of sitting the seventh grade ENLACE exam. The sample is our main estimation sample augmented with individuals who are known to have attended secondary school in the seventh grade but for whom no score is observed. The dependent variable is a binary variable that equals one if the student sits the seventh grade ENLACE exam and zero otherwise. Tele equals one if the student is enrolled in a telesecundaria and zero otherwise. Sixth grade math and Spanish scores on the ENLACE exams are measured in standard deviations. Controls include age, number of siblings, sex, prospera status, family income, mother's education, number of books in the home, and an indicator for whether the family has access to a computer. Standard errors are calculated via 250 bootstrap replications.

# B. Additional Tables and Figures

Table B-1
Instrumental Variables Regressions: First Stage

| | *Specification:* | | |
|---|---|---|---|
| | (Linear) | (Quadratic) | (Cubic) |
| Relative Distance | $-0.041^{***}$ (0.0002) | $-0.075^{***}$ (0.001) | $-0.111^{***}$ (0.002) |
| Relative Distance (Quadratic) | | $0.003^{***}$ (0.0001) | $0.010^{***}$ (0.0004) |
| Relative Distance (Cubic) | | | $-0.0003^{***}$ (0.00002) |
| Math Score (6th Grade) | $-0.00004^{***}$ (0.00001) | $-0.00003^{***}$ (0.00001) | $-0.00003^{***}$ (0.00001) |
| Spanish Score (6th Grade) | $-0.0001^{***}$ (0.00001) | $-0.0001^{***}$ (0.00001) | $-0.0001^{***}$ (0.00001) |
| Mean Score: nearest Traditional | $0.0001^{***}$ (0.00002) | $0.0001^{***}$ (0.00002) | $0.0001^{***}$ (0.00002) |
| Mean Score: nearest Telesecundaria | $0.0003^{***}$ (0.00002) | $0.0003^{***}$ (0.00002) | $0.0003^{***}$ (0.00002) |
| Age | $0.026^{***}$ (0.001) | $0.026^{***}$ (0.001) | $0.025^{***}$ (0.001) |
| Siblings | $0.009^{***}$ (0.001) | $0.008^{***}$ (0.001) | $0.008^{***}$ (0.001) |
| Female | $0.004^{**}$ (0.002) | $0.004^{**}$ (0.002) | $0.004^{***}$ (0.002) |
| Prospera | $0.065^{***}$ (0.003) | $0.063^{***}$ (0.003) | $0.061^{***}$ (0.003) |
| Family Income : Low | $-0.030^{***}$ (0.002) | $-0.029^{***}$ (0.002) | $-0.029^{***}$ (0.002) |
| Family Income : Medium | $-0.031^{***}$ (0.002) | $-0.029^{***}$ (0.002) | $-0.028^{***}$ (0.002) |
| Family Income : High | $-0.024^{***}$ (0.003) | $-0.021^{***}$ (0.003) | $-0.019^{***}$ (0.003) |
| Mother's Education : Middle | $-0.024^{***}$ (0.002) | $-0.023^{***}$ (0.002) | $-0.022^{***}$ (0.002) |
| Mother's Education : Secondary | $-0.029^{***}$ (0.002) | $-0.025^{***}$ (0.002) | $-0.023^{***}$ (0.002) |
| Mother's Education : Post-Secondary | $-0.014^{***}$ (0.003) | $-0.013^{***}$ (0.003) | $-0.011^{***}$ (0.003) |
| Books in the Home : 20 | $-0.021^{***}$ (0.002) | $-0.019^{***}$ (0.002) | $-0.018^{***}$ (0.002) |
| Books in the Home : 50 | $-0.026^{***}$ (0.002) | $-0.024^{***}$ (0.002) | $-0.022^{***}$ (0.002) |
| Books in the Home : $geq$100 | $-0.022^{***}$ (0.003) | $-0.019^{***}$ (0.003) | $-0.018^{***}$ (0.003) |
| Computer | $-0.020^{***}$ (0.002) | $-0.017^{***}$ (0.002) | $-0.016^{***}$ (0.002) |
| Rural Residence | $0.186^{***}$ (0.004) | $0.150^{***}$ (0.004) | $0.134^{***}$ (0.004) |
| Northern State | $0.056^{***}$ (0.002) | $0.068^{***}$ (0.002) | $0.052^{***}$ (0.002) |
| Speaks Language Other than Spanish at Home | 0.002 (0.005) | 0.005 (0.004) | 0.004 (0.004) |
| Municipality : 10K-20K | $-0.016^{**}$ (0.006) | $-0.023^{***}$ (0.006) | $-0.019^{***}$ (0.006) |
| Municipality : 20K-50K | $-0.023^{***}$ (0.005) | $-0.024^{***}$ (0.005) | $-0.023^{***}$ (0.005) |
| Municipality : 50K-100K | $-0.046^{***}$ (0.005) | $-0.047^{***}$ (0.005) | $-0.045^{***}$ (0.005) |
| Municipality : 100K-500K | $-0.063^{***}$ (0.005) | $-0.066^{***}$ (0.005) | $-0.063^{***}$ (0.005) |
| Municipality : $>$ 500K | $-0.087^{***}$ (0.005) | $-0.085^{***}$ (0.005) | $-0.079^{***}$ (0.005) |
| Constant | $-0.165^{***}$ (0.022) | $-0.127^{***}$ (0.021) | $-0.106^{***}$ (0.021) |

The table displays the first stage linear regressions corresponding to the IV regressions in Tables 2 and 3. The dependent variable is a binary variable that equals one if the students attends a telesecundaria and zero otherwise. To preserve the sign, the quadratic and cubic measures of relative distance are computed as $(Z_1)^2 - (Z_0)^2$ and $(Z_1)^3 - (Z_0)^3$, where $Z_1$ is the distance to the nearest telesecundaria and $Z_0$ is the distance to the nearest traditional school. Test score variables refer to the effects of increases of 100 points (about 1 sd). The omitted category in each of Family Income, Mother's Education, Books in the Home, and Municipality is the lowest one. Computer is a binary variable that equals one if the student has access to a computer at home. The sample size for all three regression is 113,525. Standard errors are robust to heteroskedasticity. $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01

Table B-2

Summary Statistics by Secondary School Type

| | General | Technical | Telesecundaria | Private | Dropped |
|---|---|---|---|---|---|
| Proportion of Cohort | 0.418 | 0.264 | 0.177 | 0.085 | 0.057 |
| Mean Math (Grade 7) | 501 | 503 | 495 | 571 | - |
| Mean Spanish (Grade 7) | 503 | 504 | 485 | 576 | - |
| Mean Math (Grade 6) | 530 | 532 | 484 | 607 | 454 |
| Mean Spanish (Grade 6) | 527 | 526 | 476 | 612 | 451 |
| Fraction Female | 0.506 | 0.497 | 0.497 | 0.511 | 0.503 |
| Mean Age (2008) | 11.9 | 12.0 | 12.2 | 11.9 | 12.9 |
| Fraction Prospera | 0.156 | 0.195 | 0.657 | 0.008 | 0.402 |

The table displays characteristics of students for all possible choices in Grade 7: general, technical, telesecundaria, private secondary schools – as well as students who drop out. The sample size for this table is $N = 183,779$. This excludes $3,117$ students who are unobserved in grade 7, but who enroll in later grades and therefore have not dropped out. Based on information presented in this table, we consider general and technical schools as a single alternative for the purposes of estimating value added between the sixth and seventh grades.

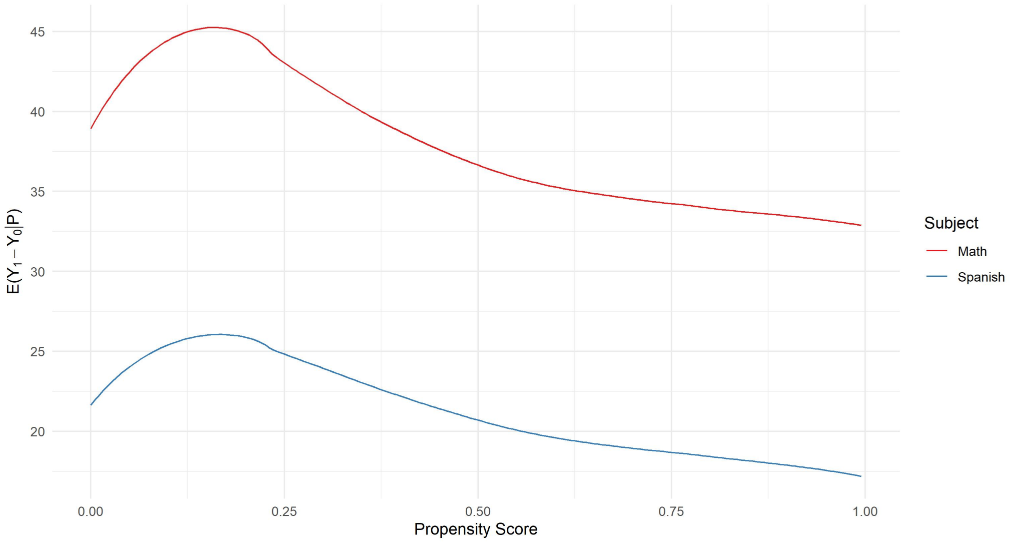
Table B-3

Variables Used in Estimation

| Variable | Definition |
|---|---|
| $Y$ | Math score in 7th grade, Spanish score in 7th grade, Indicator for progression to grade 9 |
| $X$ | **Parent:** Mother's education, family income, Prospera status, Spanish-speaking household, number of books in the home, whether the family has access to a computer.<br>**Child:** Math score in 6th grade, Spanish score in 6th grade, age, sex, number of siblings.<br>**Geography:** Rural residence, residence in northern state, indicators for six categories of municipality population, mean ENLACE score of nearest telescundaria, mean ENLACE score of nearest traditional school, mean ENLACE of secondary school attended. |
| $Z\backslash X$ | Relative distance between the nearest telesecundaria and the nearest traditional school. |

Table B-3 lists the outcome variables ($Y$), covariates ($X$), and the instrumental variable ($Z\backslash X$) that we use in our empirical analysis. Apart from the test scores variables, age, number of siblings, and the instrument, all variables are categorical. The mean ENLACE scores at each school were computed based on eighth grade scores in 2007/2008 and pertain to an older cohort than the one we study in this paper.

## Figure B-1
Sorting Based on Observables



The figure plots the observed component of the treatment effect, $X_i'(\hat{\beta}_1 - \hat{\beta}_0)$, along the y-axis as a function of the estimated propensity score along the x-axis. Higher values for the propensity score indicate a greater likelihood of attending a telesecundaria. Estimates of $\hat{\beta}_1 - \hat{\beta}_0$ are shown in Table B-4 and are obtained via the Robinson semiparametric method.

Table B-4

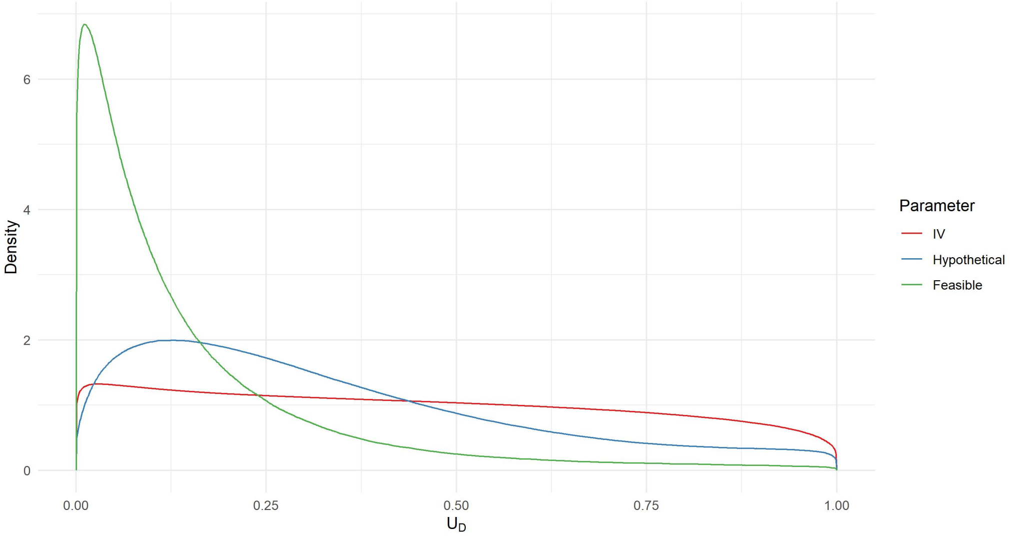
Comparison of Outcome Equation and Propensity Score Parameters

| | Effect on Attendance | $\hat{\beta}_1 - \hat{\beta}_0$ |
|---|---|---|
| Math Score (6th Grade) | -0.006 | -0.036 |
| Spanish Score (6th Grade) | -0.010 | 0.020 |
| Mean Test Score: Attended | | 0.024 |
| Mean Score: nearest Telesecundaria | 0.013 | |
| Mean Score: nearest Traditional | 0.013 | |
| Age | 0.015 | 1.63 |
| Siblings | 0.004 | 0.27 |
| Female | 0.003 | 4.66 |
| Prospera | 0.024 | -0.26 |
| Family Income : Low | -0.012 | 1.83 |
| Family Income : Medium | -0.016 | 2.33 |
| Family Income : High | -0.022 | -9.08 |
| Mother's Education : Middle | -0.016 | 6.28 |
| Mother's Education : Secondary | -0.034 | 9.47 |
| Mother's Education : Post-Secondary | -0.017 | 11.3 |
| Books in the Home : 20 | -0.011 | -1.14 |
| Books in the Home : 50 | -0.018 | -1.25 |
| Books in the Home : $\geq$100 | -0.012 | -11.0 |
| Computer | -0.021 | -4.09 |
| Rural Residence | 0.021 | -8.28 |
| Northern State | -0.029 | -11.9 |
| Speaks Language other than Spanish at Home | 0.001 | -3.74 |
| Municipality : 10K-20K | 0.000 | -7.31 |
| Municipality : 20K-50K | 0.001 | -4.47 |
| Municipality : 50K-100K | 0.002 | 3.46 |
| Municipality : 100K-500K | -0.016 | 0.29 |
| Municipality : $>$ 500K | -0.026 | 1.32 |

The table compares the average marginal effects of each variable on attending telesecundarias (from Table 4) with estimates of the treatment effect, $\beta_1 - \beta_0$, for each variable in the outcome model for seventh grade math test scores. Estimates of $\hat{\beta}_1 - \hat{\beta}_0$ are obtained via the Robinson semiparametric method.

Figure B-2

Counterfactual Treatment Parameter Weights



The figure shows the distribution of weighting functions used to construct estimates of Policy-Relevant Treatment Effects (PRTEs) for two policies discussed in section V. as well as the weights induced by Two-Stage Least Squares (IV), which uses relative distance as an instrument for telesecundaria attendance. The hypothetical policy reduces relative distance between telesecundarias and traditional secondary schools by five km, while the feasible policy constructs telesecundarias adjacent to all primary schools that do not have a telesecundaria within a five km radius. The IV weights are positive everywhere but do not correspond to either of the counterfactuals under consideration.

## C. Estimation of MTE and Treatment Effects under the Assumption of Joint Normality

The fully parametric approach to estimating the MTE specifies the unobservables in the selection and outcome equations as jointly normally distributed:

$$\begin{pmatrix} U_1 \\ U_0 \\ V \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{10} & \sigma_{1V} \\ & \sigma_0^2 & \sigma_{0V} \\ & & 1 \end{pmatrix} \right] .$$

Under these assumptions, the marginal treatment effect has the following simple functional form:

$$MTE(X, u_D) = X(\beta_1 - \beta_0) + (\sigma_{1V} - \sigma_{0V})\Phi^{-1}(U_D) .$$

We estimate the parameters $\beta_1, \beta_0, \sigma_{1V}, \sigma_{0V}$ via the two-step “Heckit" method that first estimates the propensity score via probit and then includes control functions in the outcome equations as follows:

$$\begin{aligned} \mathbb{E}(Y_1 \mid D = 1, X, Z) &= X\beta_1 + \mathbb{E}(U_1 \mid D = 1) , \\ &= X\beta_1 + \sigma_{1V} \left( -\frac{\phi(\Phi^{-1}(P(Z)))}{P(Z)} \right) , \\ \mathbb{E}(Y_0 \mid D = 0, X, Z) &= X\beta_0 + \mathbb{E}(U_0 \mid D = 0) , \\ &= X\beta_0 + \sigma_{0V} \left( \frac{\phi(\Phi^{-1}(P(Z)))}{1 - P(Z)} \right) . \end{aligned}$$

After estimating $(\beta_0, \beta_1, \sigma_{0V}, \sigma_{1V})$, we construct parametric estimates of common treatment parameters as follows:

$$\begin{aligned} ATE &= \bar{X}(\beta_1 - \beta_0) , \\ TT &= \frac{1}{N_T} \sum_{i=1}^{N_T} D_i \left[ X_i(\beta_1 - \beta_0) + (\sigma_{1V} - \sigma_{0V}) \left( -\frac{\phi(\Phi^{-1}(P(Z)))}{P(Z)} \right) \right] , \\ TUT &= \frac{1}{N_C} \sum_{i=1}^{N_C} (1 - D_i) \left[ X_i(\beta_1 - \beta_0) + (\sigma_{1V} - \sigma_{0V}) \left( \frac{\phi(\Phi^{-1}(P(Z)))}{1 - P(Z)} \right) \right] , \end{aligned}$$

where $D_i = 1$ if individual $i$ attends a telesecundaria and $0$ otherwise, and $N_T$ ($N_C$) denotes the number of students attending telesecundarias (traditional schools). Standard errors for these treatment parameters are estimated via nonparametric bootstrap.

## D. Sensitivity Analysis

We now consider alternative specifications, including those designed to investigate the validity of Assumption (A-2)′ in the main text. Our main analysis omits from the sample all children who attend a secondary school more than twenty kilometers from their primary school to eliminate the threat of endogeneity caused by families who may move to be closer to a specific school based on partial knowledge of $(U_1, U_0)$.[20]

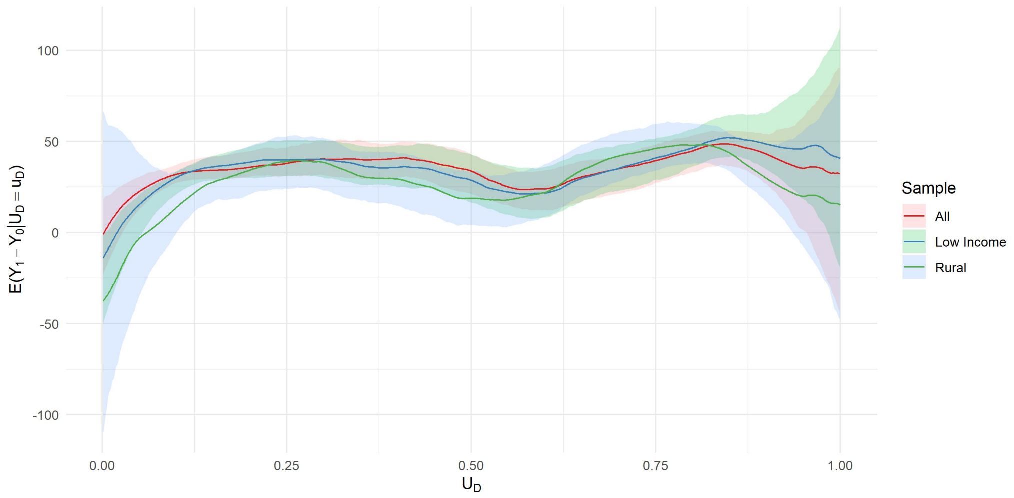
To further alleviate this concern, we replicate our analysis on subsamples composed of individuals who are unlikely to change residence for the purpose of their children attending a different school. These subsamples are the set of children living in rural areas and the set of children whose parents earn under $2500$ pesos per month (approximately $220$ USD in 2008). Figure D-1 compares the estimated MTE functions for seventh grade ENLACE scores in math with the MTE function for the entire sample. Figure D-2 repeats the analysis for seventh grade ENLACE scores in Spanish. The math MTE functions are very similar across the rural, low-income, and full samples, indicating that the pattern of sorting into telesecundarias based on unobservables that influence math scores are similar for all these populations. This suggests that assumption (A-2)′, which renders the MTE function additively separable in $X$ and $U_D$ and restricts the slope of the MTE function in $U_D$ to be the same regardless of $X$, is not overly restrictive. The MTE functions for Spanish, in Figure, D-2 instead have different behavior in the tails when conditioning on the rural and low-income subsamples, although these differences are not statistically significant. Recall that there was less evidence in Table 3 of sorting into telesecundarias on the basis of unobserved determinants of Spanish scores, so we interpret these differences as potentially resulting from sampling variation and higher variance of the semiparametric estimator.

If students must walk to school, as is common in many parts of Mexico, an additional concern may be that traveling long distances to school causes academic performance to suffer. To alleviate concerns that the instrument is directly correlated with academic outcomes, we augment our outcome equations, (1) and (2),

[20] Evidence of behavior of this sort has been used to criticize the validity of distance to college as an instrument in the United States (Cameron and Taber 2004, Carneiro and Heckman 2002), although it is less of a concern in Mexico. The source of funding for secondary schools in Mexico – 79% from the federal government with the remainder coming from state governments – reduces the incentive for hyper-local sorting driven by differences in school budgets (OECD 2019).

Figure D-1

Marginal Treatment Effect by Subsample: Math



The figure compares semiparametric MTEs estimated on two subsamples of students unlikely to move location prior to secondary school - those who grow up in rural areas and those with low-income parents ($<$ 2500 Pesos/mo). The outcome variable is seventh grade math ENLACE scores. Details regarding the estimation of the MTE functions are provided in Section IV-D. All MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

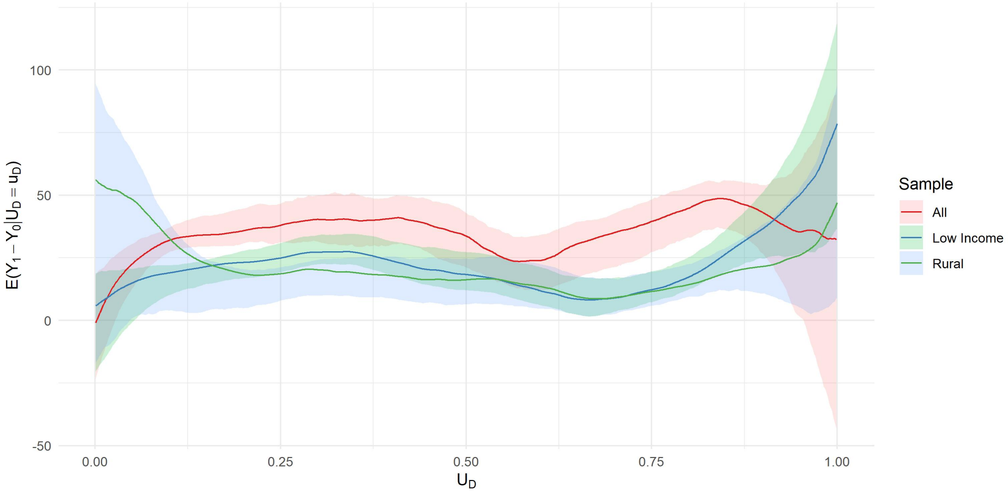
with a measure of the distance actually traveled to secondary school as an explanatory variable.

We find that including distance traveled to secondary school makes little difference in the estimated MTEs or treatment effects for math and Spanish. Figures D-3 and D-4 show that the estimated MTE curves for these specifications are very similar to those in our main analysis. The similar point estimates of treatment parameters (Tables D-1 and D-2) also reassure that our estimates of the causal effects of telesecundarias on learning are little affected by any correlation between the distance traveled to secondary school and unobserved determinants of test scores.

Throughout the paper we have measured distance "as the crow flies." We therefore conduct a final robustness check with a new instrument that is the relative distance between the nearest traditional and telesecundaria schools where distance is measured along roads and paths by Google maps. In most cases, the use of this new school measure does not cause a change in the identity of the nearest school of

Figure D-2

Marginal Treatment Effect by Subsample: Spanish

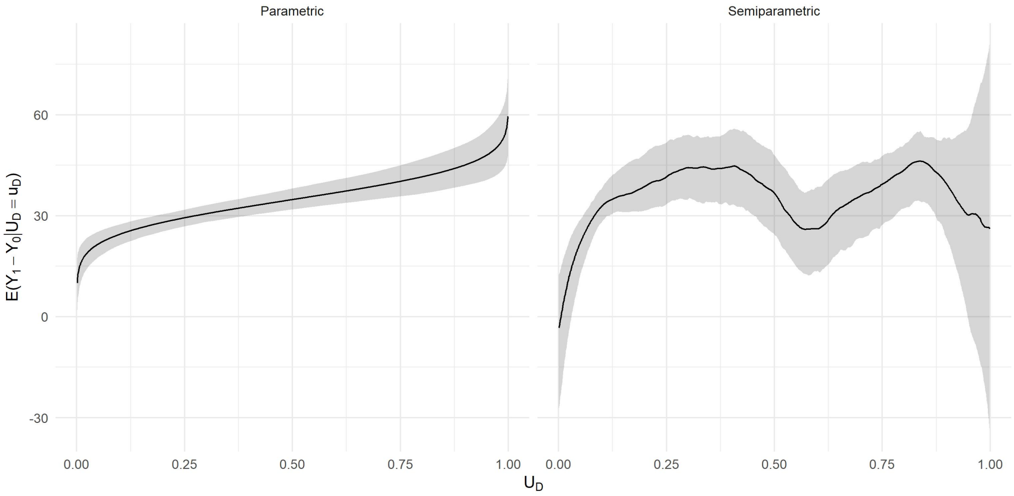


The figure compares semiparametric MTEs estimated on two subsamples of students unlikely to move location prior to secondary school - those who grow up in rural areas and those with low-income parents ($< 2500$ Pesos/mo). The outcome variable is seventh grade Spanish ENLACE scores. Details regarding the estimation of the MTE functions are provided in Section IV-D. All MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

each type, but, for a small percentage of students, the nearest school as measured by "traveled distance" differs from the nearest school measured "as the crow flies." Nevertheless, the estimates of the MTE curve and treatment parameters are little changed when we use this alternative measure of "traveled distance." We display the MTE curves for this specification in Figures D-5 and D-6, and the treatment parameters in Tables D-3 and D-4. The MTE curves and the estimated parametric and semiparametric treatment effects for math do not differ significantly from our main estimates. The Spanish MTE curve increases sharply, but noisily, as $U_D$ approaches $1$, which causes some increase in the estimates of ATE and TUT, although these differences are not more than two standard errors from the estimates in the main text.

Figure D-3

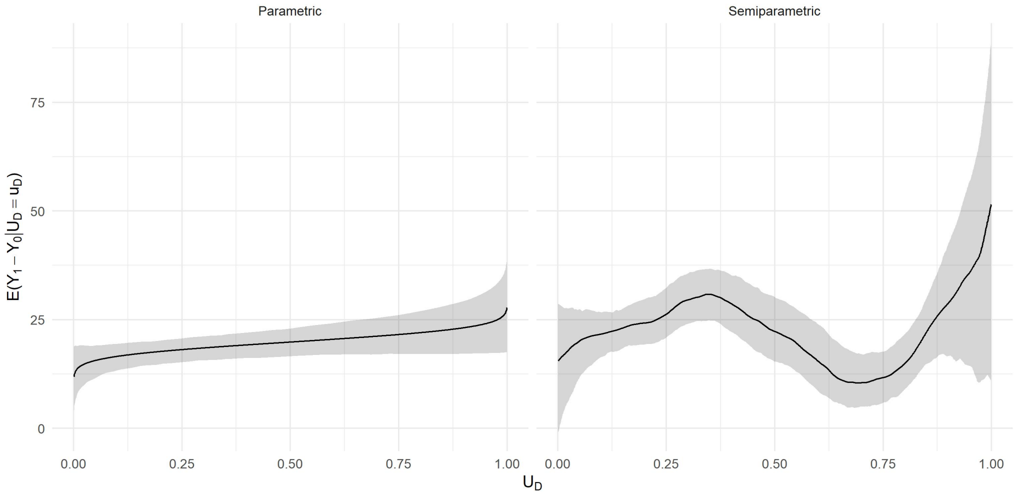
Marginal Treatment Effect Controlling for Distance Traveled: Math



The dependent variable in the outcome equation is the raw score on the seventh grade ENLACE math exam. The outcome equations include controls for sixth grade math and Spanish scores, age, sex, number of siblings, mother's education, family income, number of books in the home, family access to a computer, Prospera status, rural residence, residence in a Northern state, whether the family speaks Spanish at home, a measure of school quality, dummies for municipality size, and the distance actually traveled to secondary school. The school choice model includes the same controls and also includes the relative distance between the nearest telesecundaria and nearest traditional secondary school as an exclusion restriction. The school choice model is estimated via probit. The parametric MTE is estimated using a two-step "Heckit" procedure. The semiparametric MTE is estimated using Local Quadratic Regression and an Epanechnikov kernel with a bandwidth of $0.231$. Both MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

Figure D-4

Marginal Treatment Effect Controlling for Distance Traveled: Spanish



The dependent variable in the outcome equation is the raw score on the seventh grade ENLACE Spanish exam. The outcome equations include controls for sixth grade math and Spanish scores, age, sex, number of siblings, mother's education, family income, number of books in the home, family access to a computer, Prospera status, rural residence, residence in a Northern state, whether the family speaks Spanish at home, a measure of school quality, dummies for municipality size, and the distance actually traveled to secondary school. The school choice model includes the same controls and also includes the relative distance between the nearest telesecundaria and nearest traditional secondary school as an exclusion restriction. The school choice model is estimated via probit. The parametric MTE is estimated using a two-step "Heckit" procedure. The semiparametric MTE is estimated using Local Quadratic Regression and an Epanechnikov kernel with a bandwidth of $0.323$. Both MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

Table D-1

Estimated Treatment Effects Controlling for Distance Traveled: Math

| | Parametric | | Semiparametric | |
|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error |
| Average Treatment Effect | 34.8 | (1.97) | 35.8 | (2.78) |
| Treatment on the Treated | 30.0 | (1.53) | 26.3 | (2.12) |
| Treatment on the Untreated | 35.9 | (2.13) | 37.8 | (3.17) |

The table displays three treatment parameters corresponding to the effect of telesecundaria attendance on raw scores on the seventh grade math ENLACE exam. It differs from Table 7 in the main text in that this specification conditions on the distance actually traveled to secondary school in the outcome equations. The three treatment parameters are obtained by integrating the MTE with respect to the densities displayed in Figure 3. The simulation method of Carneiro, Lokshin, and Umapathi (2017) is used to integrate the semiparametric MTE. Standard errors are obtained by 250 bootstrap replications.

Table D-2

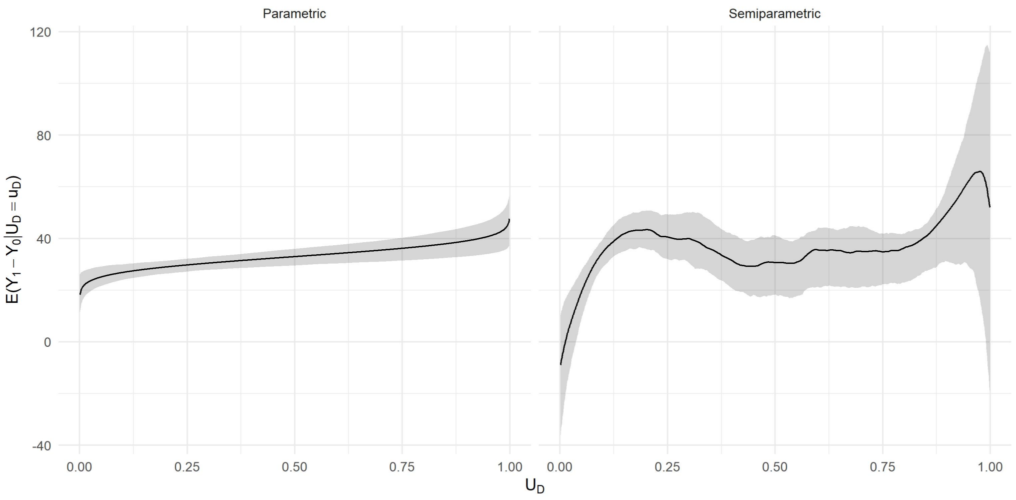
Estimated Treatment Effects Controlling for Distance Traveled: Spanish

| | Parametric | | Semiparametric | |
|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error |
| Average Treatment Effect | 19.8 | (1.93) | 22.7 | (2.65) |
| Treatment on the Treated | 18.3 | (1.58) | 18.4 | (1.96) |
| Treatment on the Untreated | 20.2 | (2.07) | 23.6 | (3.01) |

The table displays three treatment parameters corresponding to the effect of telesecundaria attendance on raw scores on the seventh grade Spanish ENLACE exam. It differs from Table 8 in the main text in that this specification conditions on the distance actually traveled to secondary school in the outcome equations. The three treatment parameters are obtained by integrating the MTE with respect to the densities displayed in Figure 3. The simulation method of Carneiro, Lokshin, and Umapathi (2017) is used to integrate the semiparametric MTE. Standard errors are obtained by 250 bootstrap replications.

## Figure D-5

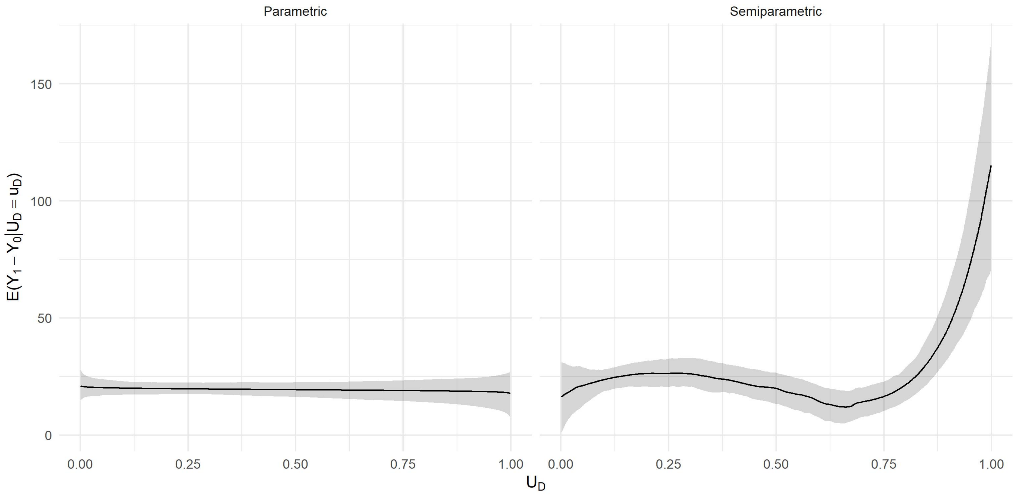
## Marginal Treatment Effect with “Traveled Distance” IV: Math



The dependent variable in the outcome equation is the raw score on the seventh grade ENLACE math exam. The outcome equations include controls for sixth grade math and Spanish scores, age, sex, number of siblings, mother’s education, family income, number of books in the home, family access to a computer, Prospera status, rural residence, residence in a Northern state, whether the family speaks Spanish at home, a measure of school quality, and dummies for municipality size. The school choice model includes the same controls and also includes the relative distance *as measured along roads and paths* between the nearest telesecundaria and nearest traditional secondary school as an exclusion restriction. The school choice model is estimated via probit. The parametric MTE is estimated using a two-step “Heckit" procedure. The semiparametric MTE is estimated using Local Quadratic Regression and an Epanechnikov kernel with a bandwidth of 0.231. Both MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

Figure D-6

Marginal Treatment Effect with "Traveled Distance" IV: Spanish



The dependent variable in the outcome equation is the raw score on the seventh grade ENLACE Spanish exam. The outcome equations include controls for sixth grade math and Spanish scores, age, sex, number of siblings, mother's education, family income, number of books in the home, family access to a computer, Prospera status, rural residence, residence in a Northern state, whether the family speaks Spanish at home, a measure of school quality, and dummies for municipality size. The school choice model includes the same controls and also includes the relative distance *as measured along roads and paths* between the nearest telesecundaria and nearest traditional secondary school as an exclusion restriction. The school choice model is estimated via probit. The parametric MTE is estimated using a two-step "Heckit" procedure. The semiparametric MTE is estimated using Local Quadratic Regression and an Epanechnikov kernel with a bandwidth of $0.323$. Both MTEs are evaluated at the mean value of the covariates, $X = \overline{x}$. Ninety percent confidence intervals are computed via nonparametric bootstrap with 250 draws.

Table D-3

Estimated Treatment Effects with “Traveled Distance” IV: Math

| | Parametric | | Semiparametric | |
|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error |
| Average Treatment Effect | 33.0 | (1.85) | 36.9 | (2.76) |
| Treatment on the Treated | 30.1 | (1.52) | 26.4 | (2.21) |
| Treatment on the Untreated | 33.6 | (1.97) | 38.9 | (3.10) |

The table displays three treatment parameters corresponding to the effect of telesecundaria attendance on raw scores on the seventh grade math ENLACE exam. It differs from Table 7 in the main text in that this relative distance measure is computed by differencing the distances *measured along roads and paths* to the nearest secondary school of each type. The three treatment parameters are obtained by integrating the MTE with respect to the densities displayed in Figure 3. The simulation method of Carneiro, Lokshin, and Umapathi (2017) is used to integrate the semiparametric MTE. Standard errors are obtained by 250 bootstrap replications.

Table D-4

Estimated Treatment Effects with “Traveled Distance” IV: Spanish

| | Parametric | | Semiparametric | |
|---|---|---|---|---|
| | Estimate | Standard Error | Estimate | Standard Error |
| Average Treatment Effect | 19.4 | (1.91) | 27.3 | (2.84) |
| Treatment on the Treated | 19.7 | (1.54) | 19.9 | (1.88) |
| Treatment on the Untreated | 19.3 | (2.04) | 28.8 | (3.22) |

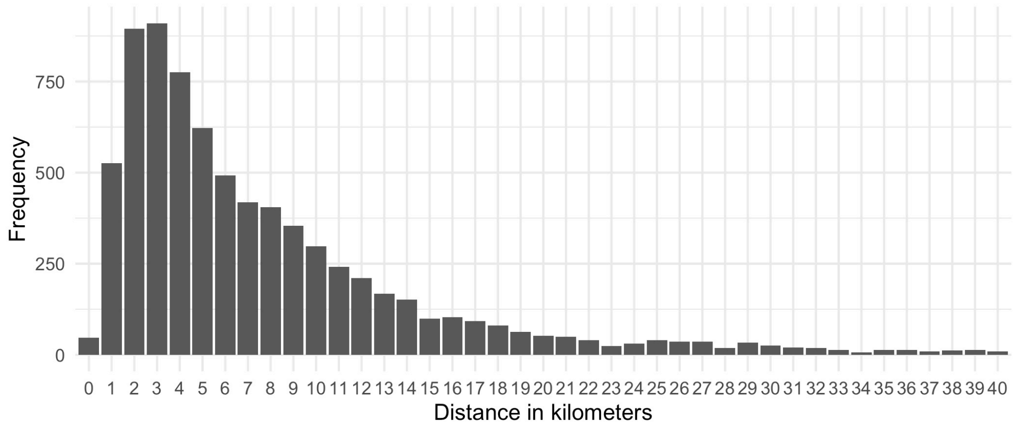
The table displays three treatment parameters corresponding to the effect of telesecundaria attendance on raw scores on the seventh grade Spanish ENLACE exam. It differs from Table 8 in the main text in that this relative distance measure is computed by differencing the distances *measured along roads and paths* to the nearest secondary school of each type. The three treatment parameters are obtained by integrating the MTE with respect to the densities displayed in Figure 3. The simulation method of Carneiro, Lokshin, and Umapathi (2017) is used to integrate the semiparametric MTE. Standard errors are obtained by 250 bootstrap replications.

## E. School Availability

Our study relies on students facing a choice between two types of secondary schools: a traditional school and a telesecundaria. In this appendix, we show that while not all students have a school of each type within a reasonable distance, a substantial fraction do, and the availability of both schools is not geographically concentrated. Second, we show that over half of students do select their nearest school (either traditional or telesecundaria) and that the students who do not choose the nearest school are not traveling that much further on average.

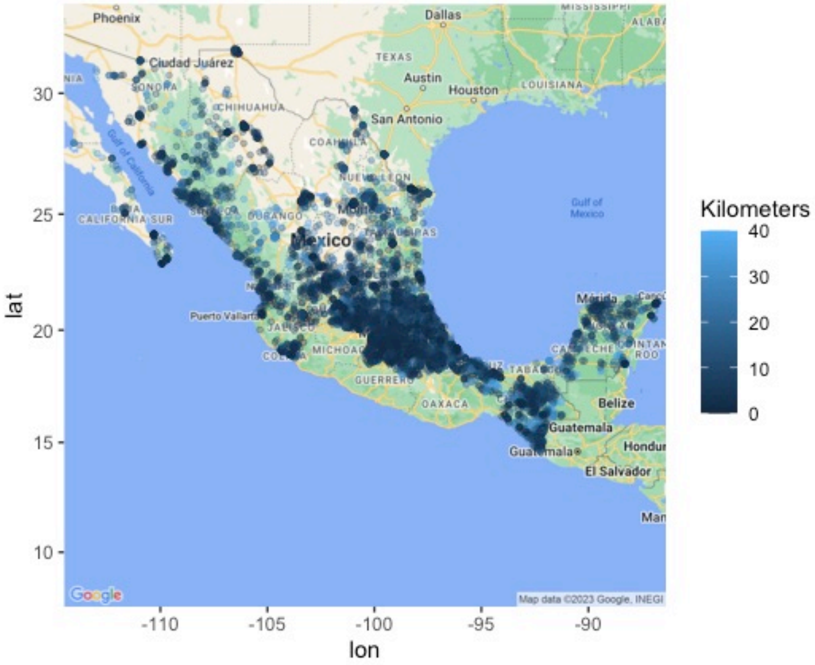
In our data, we have information on 7620 primary schools. Several primary schools do not have a telesecundaria or traditional school anywhere in their vicinity. Figure E-1 shows the maximum distance students from the primary would have to travel to reach both a traditional and telesecundaria school. This histogram only contains 7465 schools, as the top 2% of the distribution was truncated to reduce the length of the right tail. For 91.9% of the primary schools, there is a school of each type within 20 km.

Figure E-1
Maximum Distance to Reach Both a Traditional and Telesecundaria from Each Primary



We next investigate whether the local availability of the two school types varies across regions in Mexico. Figure E-2 is a map of the location of the primary schools in the full data set, once again removing those in the top 2% of distances. The color of the school indicates how far students would have to travel to reach a school of each type. There is good coverage over the various regions and main population centers in Mexico. As mentioned in the main text, we do not include primary schools from Guerrero, Oaxaca and Michoacan due to reporting concerns. Figure E-3 uses the same methods but plots only the schools that are in our estimation sample (6824 schools). The extent of coverage is not much affected by subsetting to these schools.

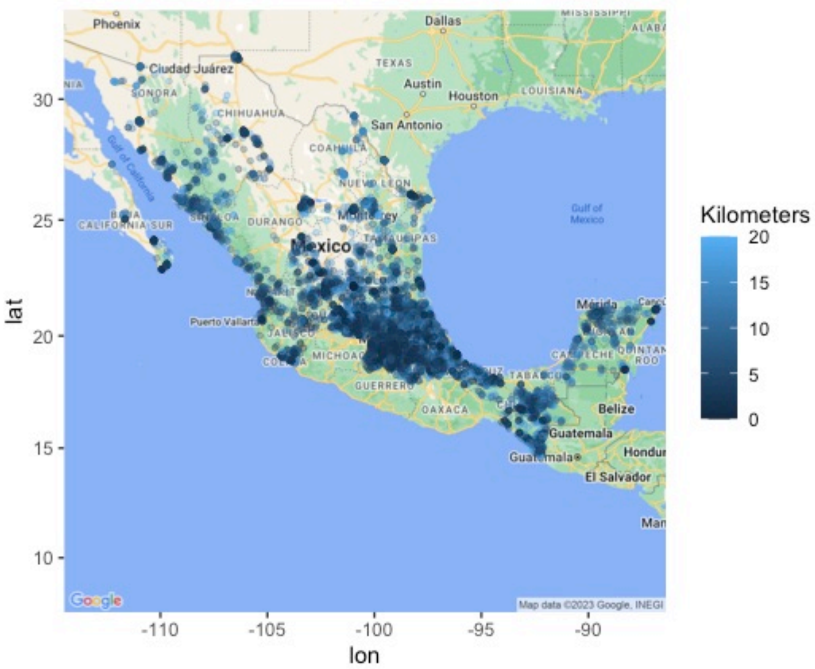
Figure E-2
Geographic Location of All Primaries in Data



Next, we consider the fraction of students that select their nearest school. Overall, 54.1% of students attend either their nearest telesecundaria or nearest traditional school. For those that select into telesecundarias, this group is even higher (86.1% of students who end up attending a telesecundaria attend the closest telesecundaria). For students who end up attending a traditional school, 45.8% attend the nearest one. The lower value for traditional schools is in part due to the fact that there are two types of schools grouped in this category: general and technical schools. For our analysis, these are grouped together as the outside option. For students who attended a telesecundaria, but not the closest one, they traveled an extra 4.2 km on average. For students who attended a traditional school, but not the closest one, they traveled an extra 2.0 km on average.

Figure E-3

Geographic Location of All Primaries in Estimation Sample



## F. Sample Selection

This section outlines the steps taken to construct our estimation sample. There were over $1.94$ million 6th grade students who wrote the ENLACE in the 2007/2008 school year. The following table explains the steps we took to go from the universe of Mexican schoolchildren to our estimation sample. The largest decrease in sample size stemmed from merging in data from surveys that were administered to all students in a random sample of primary schools.

| Sample Size | % Decrease | Notes |
|---|---|---|
| 1,942,508 | - | All 6th grade students who wrote the ENLACE in the 2007/2008 school year. |
| 1,796,745 | 7.5 % | Dropped students in the states of Guerrero, Michoacan and Oaxaca due to incomplete reporting. |
| 1,770,949 | 1.4 % | Dropped students who switched states between Grade 6 and 7, since we cannot accurately predict their choice sets. |
| 1,730,835 | 2.3 % | Dropped students who have non-traditional grade progressions (student is retained or skipped a grade, or there was a coding error). |
| 1,713,770 | 1.0 % | Dropped students who attended a primary school for which we could not find geographic coordinates. |
| 1,706,597 | 0.4 % | Dropped students who attended a secondary school for which we could not find geographic coordinates. |
| 202,972 | 88.1 % | The sample was merged with the students who also filled out the student survey (these students attended a random sample of primary schools). |
| 186,896 | 8.0 % | The sample was merged with the students whose parents also filled out the parent survey. |
| 173,735 | 7.0 % | Drop students who we do not see enrolled in 7th grade. |
| 173,322 | 0.2 % | Drop students who have non-standard school type codes. |
| 171,413 | 1.1 % | Drop students with high likelihood of copying on exam. |
| 155,909 | 9.0 % | Drop students who attended a private secondary school. |
| 139,424 | 10.6 % | Drop students who attend a secondary school more than 20km from their primary school. |
| 122,195 | 12.4 % | Drop those missing any variable in $X$ |
| 118,526 | 3.0 % | Drop students whose relative distance measure is greater than 20km. |
| 113,525 | 4.2 % | Impose common support for the propensity score. |

## G. Measuring School Quality

A potential concern regarding regarding instrumental validity is that relative distance could be correlated with the difference in quality between local telesecundaria and traditional secondary schools. In this section, we use test scores of older students to proxy for the quality of secondary schools. We construct a quality measure for each secondary school by averaging the combined math and Spanish scores for students who were in the eighth grade in 2009, who are one cohort older than students in our estimation sample. Given that these students were enrolled in the same time period, we believe that their test scores should be a reasonable proxy for the quality of the school.

Table G-1 shows regressions that examine the correlation between relative distance and school quality. The dependent variable in columns (1) and (2) is the difference in eighth grade test scores between the nearest telesecundaria and the nearest traditional secondary school. The unit of analysis for these regressions is a primary school, which explains the smaller sample size. Column (1) of Table G-1 shows that relative distance has a slight negative correlation with the difference in test scores between the two schooltypes. The coefficient is statistically significant, although not economically large, with each additional kilometer of distance being associated with a reduction in the difference in test scores between telesecundarias and traditional schools of under 1 point (which is less than 1/100 of a sd). Controlling for municipality population and a rural dummy, in column (2), reduces the size of the coefficient, and it is no longer statistically significant.

It is important to include the rural dummy and the municipality population variables, because primary schools in rural areas are closer to telesecundarias and further from traditional secondary schools compared to primary schools in urban areas. This results in relative distance typically having a larger negative value in rural areas. For this reason, the analysis in the paper has included controls for rural regions and municipality population. The main analysis also controls for school quality at the secondary school that students actually attend in the outcome equation.

Table G-1

Correlation between Instrument and School Quality

| | *Dependent variable:* | | | |
|---|---|---|---|---|
| | Difference in Quality | | Telesecundaria Quality | Traditional Quality |
| | (1) | (2) | (3) | (4) |
| Relative Distance | $-0.733^{***}$ | $-0.039$ | $-0.195^{*}$ | $-0.120$ |
| | (0.110) | (0.137) | (0.115) | (0.087) |
| Constant | $-5.79^{***}$ | $-9.08^{***}$ | $482^{***}$ | $490^{***}$ |
| | (0.704) | (1.315) | (1.101) | (0.831) |
| Rural Dummy | No | Yes | Yes | Yes |
| Municipality Pop | No | Yes | Yes | Yes |
| Observations | 6,854 | 6,248 | 6,281 | 6,342 |
| $R^2$ | 0.006 | 0.028 | 0.015 | 0.012 |
| Adjusted $R^2$ | 0.006 | 0.027 | 0.014 | 0.011 |

The table shows OLS regressions of school quality measures on relative distance. School quality for each secondary school is measured by the average combined math and Spanish ENLACE scores for students who were in the 8th grade in 2009, one cohort older than students in our main estimation sample. The unit of analysis is a primary school. The dependent variable in columns (1) and (2) is the difference in school quality between the nearest telesecundaria and the nearest traditional secondary school. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.