FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE
DEPARTAMENTO DE ECONOMÍA

SERIE DE DOCUMENTOS DE TRABAJO

SDT 472

The impact of commuting time over educational achievement: A machine learning approach

Autores:
Dante Contreras
Daniel Hojman
Manuel Matas
Patricio Rodríguez
Nicolás Suárez

Santiago, Noviembre de 2018

sdt@econ.uchile.cl
econ.uchile.cl/publicaciones

The impact of commuting time over educational achievement: A machine learning approach[*]

Dante Contreras[†] Daniel Hojman[†] Manuel Matas[‡] Patricio Rodríguez[‡] Nicolás Suárez[†]

November 29, 2018

Abstract

Taking advantage of georeferenced data from Chilean students, we estimate the impact of commuting time over academic achievement. As the commuting time is an endogenous variable, we use instrumental variables and fixed effects at school level to overcome this problem. Also, as we don't know which mode of transport the students use, we complement our analysis using machine learning methods to predict the transportation mode. Our findings suggest that the commuting time has a negative effect over academic performance, but this effect is not always significant.

Keywords: Student commuting, Scholar achievement, Machine learning.

JEL Codes: C26, I29, R23

[*] We thank professors Alejandra Mizala, Valentina Paredes, Jaime Ruiz-Tagle, Sergio Urzúa and the assistants to the internal seminars of the Department of Economics for their valuable comments. Nicolás Suárez thanks professors Dante Contreras and Daniel Hojman for their guidance in his thesis project, and he acknowledges the financial support provided by the National Commission of Research in Science and Technology (CONICYT). National Master's Scholarship 2017-22171231.

[†] Department of Economics, Faculty of Economics and Business, Universidad de Chile.

[‡] Center for Advanced Research in Education (CIAE), Universidad de Chile.

1 Introduction

It's no secret that education have several positive effects over society: Several authors have stated that education is an important determinant of economic development[1] (Colclough, 1982; Temple, 2002; Hanushek and Wößmann, 2007; Becker and Woessmann, 2009; Ciccone and Papaioannou, 2009). Other articles have found a link between education and crime reduction (Lochner and Moretti, 2004; Machin et al., 2011), mother's education and children health (Currie and Moretti, 2003), education and civic participation (Dee, 2004; Milligan et al. 2004), education and social capital (Temple, 2001), among other possible effects of education. Considering this, it's clear why education is a mayor concern for policymakers, especially in developing countries.

As stated in Rivkin, Hanushek and Kain (2005), formal education is a function of several factors, including school experiences, community, family and others. Improvements only in school attainment may not imply economic development, because the quality of the supplied education is a key factor in development (Hanushek and Wößmann, 2007). This shows us that education is a complex interaction of several factors, and because of that, there are various studies focusing on different variables that affect the outcomes of the educational systems: Some authors have studied the effect of the classroom size on the students achievement (Angrist and Lavy, 1999; Krueger, 2003), the effect of the match between the student and the teacher gender (Dee, 2007; Fryer and Levitt,2010; Paredes, 2014), teacher absenteeism (Duflo and Hanna 2005; Banerjee and Duflo, 2006), among other factors.

A key factor that have received little attention by investigators is the commuting time, understood as the time students expend traveling from home to school and vice versa. The time the students spend traveling between home and school could be alternatively used studying (which could help them in their learning process), sleeping or relaxing with recreational activities, which could improve their psychological welfare. Because of that, commuting time have an effect not only over the learning process, but also over the welfare of the students. Studying the effect of commuting time over students is specially relevant in cities like Santiago, where the average commuting time of workers is 53 minutes, and the average commuting time of students is 34 minutes[2].

[1] Some papers measure economic development as gdp growth, and others as value added and employment growth.

[2] Own calculations based on the 2012 Origin and Destination Survey.

Some articles have studied the link between commuting time and achievement around the world: In Sweden, Kjellström and Regnér (1999) argue that the distance to the nearest university have a negative impact on the likelihood of starting higher education, and according to Westman et al. (2015), both commuting time and travel mode have an impact on the cognitive performance and current mood of students. In England, Gibbons and Vignoles (2012) found that geographical distance has little or no impact on the decision to participate in higher education, but has a strong influence on institutional choice, while Dickerson and McIntosh (2013) found that the distance from the student's home to the academic institutions have a negative impact on their probability of studying more than the compulsory time. In Netherlands, Sá et al. (2006) showed that geographical proximity increases the odds of high school graduates continuing their education at a university or professional college, while Kobus et al. (2015) sustained that the university students with greater commuting times visit their universities less often, and when they visit it, they stay longer than other students, and have worse academic performance. Other articles support that higher commuting times reduce the graduation rates in students from Norway (Falch et al. 2013), and that the commuting time have a negative effect on the academic performance of sixth-grade students from Brazil (Tigre et al., 2017).

In Chile, there is no direct evidence concerning the impact of commuting time over achievement. A related article is Asahi (2016), which tries to estimate the impact of the expansion of the subway network in Santiago on the student's spatial mobility and their school's performance. The author finds that when the school become more accessible because of the network expansion, the enrollment in the near schools increases, but also the performance of their students diminish. Other authors studied the relationship between school choice and distance to school: Gallego and Hernando (2008) found that 2 of the attributes that the parents value the most when choosing a school are the school standardized test's results and the distance to the school, and that parents face a trade-off when choosing between them. Chumacero et al. (2011) confirm their findings, using a dataset with much more accurate data about the distance from the house to the school.

Because of this lack of evidence, in this article we will study the causal effect of the commuting time over academic achievement with a sample of 8th graders from Santiago. Specifically, we will use the 8th grade

SIMCE dataset from 2013, and commuting time data provided by the Center of Advanced Research in Education (CIAE). As we know the distance to the school and the commuting time in every transportation mode for each student, but we don't know the in which mode of transport the student commutes, we need a way to predict their transportation mode choice. For this purpose, we will estimate a mode of transport prediction model, using the 2012 Origin and Destination Survey, a dataset that contains information about the mode of transport used by the students. When this model is calibrated, we can predict the transportation mode in our SIMCE dataset, and then we can generate some commuting time measures.

We have described our general approach to the mode of transport prediction problem, but now we have to find an appropriate prediction model for this problem. A possible solution is to apply machine learning techniques: As stated in Athey (2017) and Mullainathan and Spiess (2017), supervised machine learning methods are a great contribution to the field of economics, as they allow us to estimate complex prediction models with high accuracy. One of the drawbacks of machine learning models is that we can't interpret or make inference about the parameters of the model, but in this case this is not a problem, because it is not our main interest to understand how families choose the mode of transportation. Because of these reasons, machine learning techniques are the most suitable to solve this prediction problem.

The Chilean educational system and our particular dataset have some peculiarities that make the estimation of the causal effect of commuting time over academic achievement more challenging than in other articles: For instance, in Tigre et al. 2017, they work with a sample of 118 public schools in the city of Recife, Brazil, with 2,483 6th grade students, where the correlation between commuting time and performance is negative. Less than 2% of the parents in their dataset reported choosing the location of their home based on its proximity to schools, and parents are encouraged to enroll their children in nearby schools. When they analyze the effect of commuting time over achievement, the endogeneity caused by the school and locational choice of the families is a second order problem, but nevertheless they adopt an instrumental variables approach to address potential omitted variables bias.

In our case, we have a dataset with more than 23,000 8th grade students, that assist to around 1,400 different schools, that include public, subsidized private and private schools, and where the correlation between distance to school and performance is positive, even when we calculate it separately for our 3 types of school dependences. Because of the voucher educational system implemented in Chile, the families can choose among a wide variety of schools to enroll their children, and they could also potentially choose where to live, so the commuting time will generate an endogeneity problem in our model. Because of this, we need to develop a methodology that allows us to control for this serious choice problem, and help us to discover if the effect of commuting time over achievement is really positive in our sample.

As we will argue later, the residential choice is not a problem when we talk about families in Santiago, as articles like Méndez and Goya 2018 help us to argue that most of the families don't base their housing decision on the availability of nearby schools. Also, as stated in Gallego and Hernando 2008, and in Chumacero et al. 2011, parents' school choice decision is mainly based on school quality and the distance between home and school. Because of that, we will estimate the impact of commuting time over achievement using an instrumental variables/ fixed effects approach: We will use fixed effects at school level, so we can control for common unobservable factors that all the students share in the school, related to the way their parents made the decision of enrolling them in their school, and this will certainly capture the school quality's dimension of the school choice process.

Regarding our instrumental variables, following Kobus et al. 2015 and Tigre et al. 2017, we will instrument the commuting time to the actual school with the average commuting time to the 2 nearest schools[3]: The idea of this is that, after controlling for school quality, the marginal decision of where to enroll a children must be associated to commuting time costs. We can capture these commuting time costs with our instrument, as in our first stage we will be able to compare the actual commuting time with the commuting time to nearby schools.

Our results suggests that, if we don't control for our endogeneity problems, the effect of commuting time over academic achievement appears to be positive, but when we use our instrumental variables/fixed effects methodology, we find that the effect of commuting time over academic achievement is negative, but not always significant.

This article is a contribution to the literature of the effects of commuting time: This kind of dataset have

[3] We don't consider the actual school of the student among this 2 nearest schools.

never been used in Chile, and the previous studies were focused on the preferences of the household when they make choices, so this is the first study on this field. Also, this article contributes to the literature of commuting time in a econometric sense, as we add fixed effects to the instrumental variables approach[4]. Finally, by combining georreferenced data, the Google Maps API and machine learning prediction models, we also contribute by constructing an innovative measure of commuting time.

2 Methodology

The objective of our model is to estimate and quantify how the students achievement is affected by their commuting time. The problem with our dataset is that we know the distance to the school, and the calculated commuting time[5] in every mode of transport, but we don't know in which mode the student choose to travel. Because of this, we will model the problem in the following way:

1. We estimate the effect of our measures of commuting time over the student achievement.
2. Before that, we apply machine learning methods to estimate a mode of transport model. We calibrate these models in an auxiliary dataset that contains information about the mode of transport choice for a sample of students who have similar characteristics to the individuals in our original dataset.
3. After estimating the mode of transport model, we use these results to predict the mode of transport in our original dataset. With this, we can construct the measures of commuting time that we use in step 1.

Now we will describe with more details every step of our methodology:

2.1 Academic achievement model

We will model how the student's achievement is affected by commuting time. Initially, we can say that the student i outcome, Y_i is determined by some of his individual characteristics, X_i and his commuting time to school, CT_i. This renders the following equation:

$$Y_i = CT_i'\beta + X_i'\gamma + e_i$$

[4] In Falch et al. (2013) they also use fixed effects, but not at a school level, and they don't use them to solve the endogeneity problem.

[5] By "calculated commuting time" we mean the commuting time between 2 points, which is calculated using the Google Maps API.

The problem with this equation is that from here we cannot estimate β in a consistent way, because there is a correlation between the commuting time and our error term e_i, as we can think that there are some unobserved factors related to school choice that can affect the commuting time of the students. Indeed, the ability of the families to choose the location of their home and to choose a school make the commuting time an endogenous variable. If we don't take this endogeneity into account, we would have a problem of omitting relevant variables, and this could upward bias our commuting time coefficient: because there could be parents that send their children to far away schools, but if these schools are good there are going to be gains in the academic achievement that could even offset the negative effects of the time lost traveling to the school.

There are some articles, like Ferreyra (2007), in which the housing and schooling choices are made jointly, but here we will follow Chumacero et al. (2011), and we will suppose that the location of the houses of the families is exogenous, so our endogeneity problem is restricted only to the school choice problem. Indeed, there is evidence for Chile that shows that a minority of the households base their residential decisions on the availability of nearby schools: In the third chapter of Méndez and Gayo (2018) is explained that most of the households in Santiago have no real options to choose from, some households base their residential decision on the neighborhood, and a minority of families choose based on the availability of schools (less than 10% of the families in low and middle income areas, between 10% and 20% of the families in high income areas).

To tackle the endogeneity caused by the school choice problem, we will use instrumental variables and fixed effects at school level. The following equations describe the achievement of the student i in school s:

$$CT_{is} = Z_i'\rho + X_i'\delta + \phi_s + \epsilon_{is} \tag{1}$$

$$Y_{is} = CT_{is}'\beta + X_i'\gamma + \varphi_s + \varepsilon_{is} \tag{2}$$

In our equations, Z_i is an instrumental variable, and ϕ_s and φ_s are school level fixed effects. As stated by Gallego and Hernando (2008) and Chumacero et al. (2011), families chose schools based on their quality and the distance to the schools, so the idea of our fixed effects approach is to account for common unobservable characteristics that lead households to make the same schooling choices.

The former methodology controls for school quality, but doesn't eliminates the endogeneity problem completely: After taking into account the school quality factor, the school choice should be based majorly in commuting time related costs. Here, our instrumental variable plays a role: We will use the average commuting time to the 2 nearest school to instrument our endogenous variable, the commuting time to the actual school of the student. If we look at our first stage in equation (1), we can see that now we can model the commuting time to the actual schools in terms of the commuting time to the 2 nearest school, a variable that captures the local supply of schools and helps us to measure the relative cost of traveling to a school that is near or far away. Also, in our first stage we also have fixed effects, so we are controlling for unobserved factors that influenced parents in the same school to choose a school at a determined distance from their house.

Our instrument is relevant, because it contains information about the costs associated with commuting time, and also it satisfies the exclusion restriction, because the distance to schools that the student doesn't attend to cannot affect his or her outcome.

2.2 Mode of transport prediction

We model how families decide how their children will commute between their household and the school. Thus, the students travel to school by car, public transport or walking.

We can define our dependent variable M_i, as follows:

$$M_i = \begin{cases} 1 \text{ if the student } i \text{ travels by car} \\ 2 \text{ if the student } i \text{ travels in public transport} \\ 3 \text{ if the student } i \text{ travels by foot} \end{cases}$$

We can also define T_i as a vector characteristics of the student, her family, the distance to her school and the commuting time associated with each transport alternative.

2.2.1 Machine learning models

With this dependent and independent variables, we need a model to explain the mode of transport choice. For this, we will use supervised machine learning methods: These method consist in estimating and calibrating a prediction model, which purpose is to fit well with the current data, and then work in a reliable way when facing new data. Here, 2 key concepts are **underfitting** and **overfitting**: underfitting is the situation when the model have a bad fit with the known data and is not able to learn the underlying relations hidden in the data, whereas overfitting occurs when the model memorizes the data patterns too well, and loses its capability to do good predictions when faced with new data.

Because of this, one concern when estimating machine learning model is how to calibrate a model in order to avoid underfitting and overfitting. One of the most basic approaches is to split the sample in two samples, train the model in one sample, and then predict the dependent variable in the second sample, and with that the accuracy of the prediction can be measured. Here, we will perform 2 splits of our sample:

1. First, the dataset will be split in 2 parts: One containing the 90% of the sample, to train the models, and the other containing the remaining 10%, where the accuracy of the predictions will be tested.
2. Then, the 90% of the sample will be split in 10 folds of the same size: Each model will be estimated 10 times, every time 9 folds will be used to train the model, and the remaining fold will be used to measure the accuracy of the prediction. Then, the predictions of the models are averaged to make the final prediction. This procedure is called cross-validation, and allows our model to be less sensitive to outliers in the data, because they are left-out in some estimations, so the models can learn more general patterns.

From the last paragraph, the first split of the data may seem unnecessary, but it serves a purpose: We will estimate several models, and we will tune their parameters to improve their predictive power. This tuning imply that, in a certain way, the models are learning the patterns of the left-out data, because the parameters are modified in a way that the accuracy in the left-out data is improved, thus they may be overfitting. Because of this problem, the models are estimated and tuned with the 90% dataset split in 10 folds, and the accuracy of every tuned model is calculated with the 10% left-out data. This allows us both to compute the accuracy of every model with previously unseen data, and to compare the accuracy of different models.

Finally, we can explain our procedure: First we define 7 machine learning models that are going to be applied here: K-Nearest Neighbors, Gaussian Naive Bayes, Gradient Boosting Decision Trees, Logistic regression, Neural Networks, Random Forest and Support Vector Machine. We will not explain what are these models, and it's not really necessary, because for our purpose they are only prediction models. These 7 models will be trained and tuned with our 90% sample split in 10 folds. To tune the parameters of these models, we have defined grids of parameters, and the best model was defined as the one with the higher accuracy when predicting the values of the variables in the left-out fold. Then, for the best model in every family of models, the accuracy of the prediction is computed with the 10% left-out sample, and with this, we can see which of the 7 models generates the most accurate prediction.

2.3 Commuting time prediction and measures

In the previous step, we tuned 7 machine learning models, so we can use these trained models to predict the mode of transport with new data. Thus, we predict the mode of transport of the students in our main dataset, $\hat{M_i}$, based on a vector of characteristics T_i'. If we define the calculated commuting time for the student i in the mode of transport j as CC_{ij}, we can compute the **predicted commuting time**, defined as the calculated commuting time in the predicted mode of transport. This is defined as:

$$PCT_i = \sum_{j=1}^{3} (\hat{M_i} = j)\, CC_{ij}$$

Where $(\hat{M_i} = j)$ is 1 if the condition is true, and 0 if not. Also, to prevent potential problems with our prediction model, we will also measure of commuting time as the **traveled distance**, defined as the distance that the students cover when they travel between the school and their houses. This measure is calculated as the shortest route between the 2 mentioned points, and this route have to be defined over streets[6]

We have 2 types of measures, but as the predicted commuting time can be computed with the results of our 7 machine learning models, we will have a total of 8 commuting time measures that will be used in the first step of our procedure.

[6] By this we mean that this is not a measure of linear distance, that is a way of calculating distance as a straight line between 2 points, and doesn't considers if there are streets that allow the individuals to travel that path in the real world.

3 Data

In this section, we will briefly describe our datasets: Our main source of information is the dataset that contains the results of the 2013 8th grade's SIMCE[7] tests. This dataset includes information about the student's test results in mathematics and spanish, the school to whom they attend, information about their family, among other variables. We will add to this dataset information about the past achievement of this students, so we will also use the 2009 4th grade's SIMCE tests. Generally, the investigators don't know where the student lives, but we have access to a dataset built by the CIAE and the CIT, that contains georeferenced data about the student's residence. With this data we can calculate the commuting time in every mode of transport, with the aid of the Google Maps API. For this purpose, we estimate the commuting time by car, public transport and walking, assuming that the student travel from home to school a monday at 07:00 AM.

Because of the potential problems with rural students (it's harder to georeference their homes), we will restrict our sample only to students in the "Gran Santiago"[8]. In order to show the geospatial distribution of the variables of interest, we will use the INE (National Institute of Statistics) maps of the communes of Santiago, updated up to the 2016 pre-census study. We lost a lot of data because of problems of recollection and consistency in the SIMCE datasets and the commuting time dataset. In the appendix 8.1 we will explain how much observations we lost in every stage of the building and cleaning processes, and why we think our sample is representative of the whole city.

In order to estimate the parameters of our mode of transport choice model, we will use the data from the 2012 Origin and Destination Survey, a rich dataset that contains information about the characteristics of the individuals and their families, the coordinates of the starting and ending point of their travels, their departure time, the mode they chose, and others. We will restrict our sample to travels made by individuals between the ages of 12 and 15 that have information about their parents, and that traveled for academic purposes. Here, we will also use the Google Maps API to generate the calculated commuting times.

[7] Measurement system of the quality of the education

[8] Includes the Santiago province and the adjacent communes of Puente Alto and San Bernardo.

In the table 1 we can see a brief description of some of the variables contained in our dataset. Because the SIMCE scores have been standardized, it's no surprise that their means are 0 and their standard deviations are 1. We can see that our sample contains as much men as women, and that their parents have in average almost 12 years of schooling. We can also notice also the majority of the students come from subsidized private schools, and only a few come from non-subsidized private schools.

Table 1: Descriptive statistics

Variables	**Mean**	**Std. Deviation**	**Median**	**Minimum**	**Maximum**
Mathematics score	0.00	1.0	-0.01	-2.74	2.72
Spanish score	0.00	1.0	0.08	-2.92	2.13
Average score	0.00	1.0	0.02	-3.09	2.65
Woman	50.06%	-	-	-	-
Mother years of schooling	11.73	3.10	12	0	23
Father years of schooling	11.81	3.37	12	0	23
School Dependence:					
Municipal	33.44%	-	-	-	-
Subsidized private	60.88%	-	-	-	-
Non-Subsidized private	5.66%	-	-	-	-
Observations	22,858				

In the table 2 we compare the variables in our main dataset (SIMCE scores) with our auxiliary dataset (Origin and Destination Survey), in order to see if our two samples for the prediction model are similar. We can see that, in general, the mean of the variables are similars, except with the calculed commuting time by foot, which is significantly bigger in the Origin and Destination Survey. Also, we can see that our samples are very different in their sample sizes.

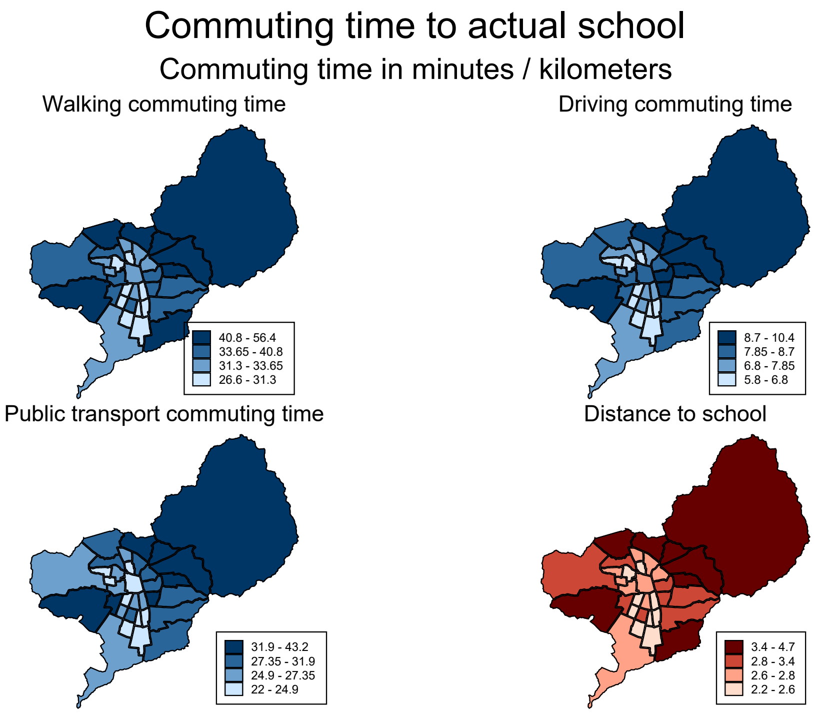
In the figure 1 we can see the geographical distribution of the calculated commuting times and the distance between the house and the school. We can see that in general, the people in the center of the city faces shorter commuting times that the people in the periphery of the city, including the people in the north-eastern zone of the city (the wealthiest zone).

Next, in order to analyze how diverse are the schools in terms of commuting time, the figure 2 plots the distribution of the interquartile range within each school, using different measures of commuting time. From the distance interquartile range we can see that the students in more than 70% of schools live approximately in a ratio of 5 kilometers around the school, we can see that the walking time distribution

Table 2: Dataset means comparison

Variables	**Dataset**	
	Main	**Auxiliary**
Woman	50.1%	49.62%
Household income	$548,829	$718,383
Mother years of schooling	11.73	10.61
Father years of schooling	11.81	11.1
Distance to school	3.2	3.85
Calculated Commuting time:		
By car	7.91	10.4
By public transport	28.9	27.55
By foot	37.1	58.75
Observations	20,858	1,969

Figure 1: Commuting time communal distribution



Commuting time to actual school

Commuting time in minutes / kilometers

Walking commuting time

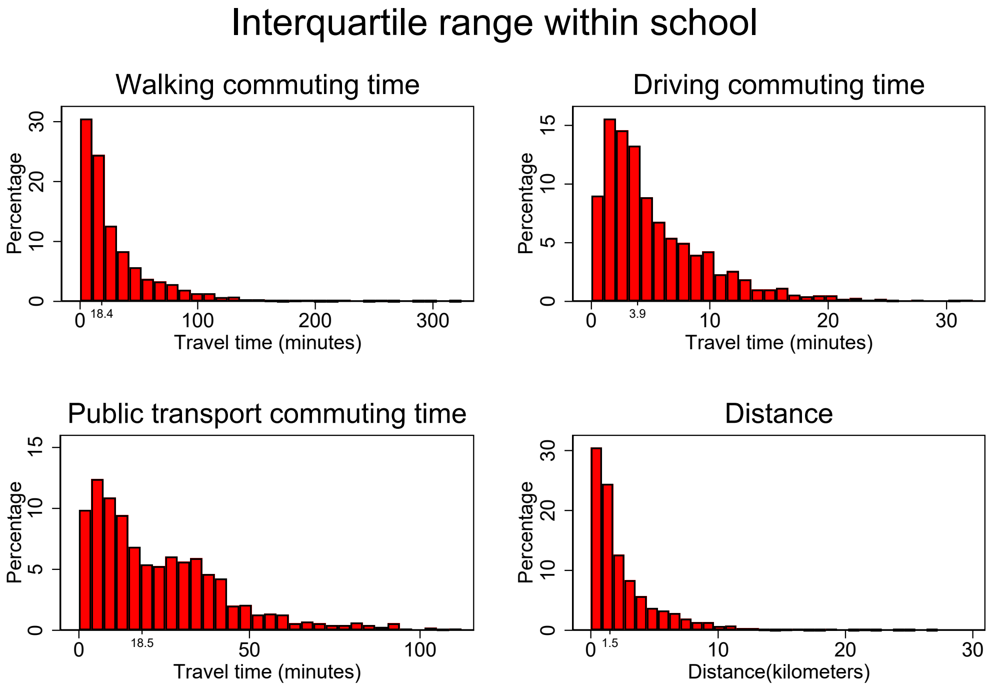
Driving commuting time

Public transport commuting time

Distance to school

Source: Own elaboration based in the SIMCE/CIAE data is almost the same as the distance distribution, and the driving is similar, except for that the first bin is smaller that the second one, likely because sometimes it's inefficient to travel by car when the distances are shorts, so in some cases the travel time associated with short distances is plotted in the second bin of the plot. Finally, the public transport distribution is less conventional, showing that because of the design of the city, even if the students live relatively near the school, their travel time can differ dramatically depending of from where they depart.

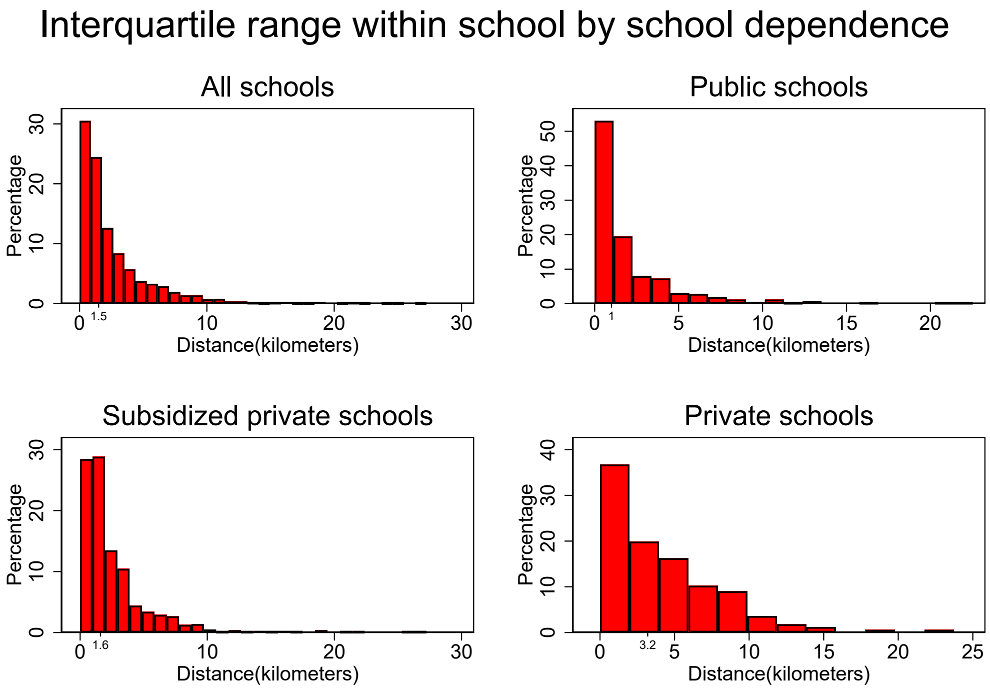
Figure 2: Differences within school in commuting time



Source: Own elaboration based in the SIMCE/CIAE data

Finally, the figure 3 shows us the distribution of interquartile range of the distance to the school, now separated by school dependence. This figure shows us that, generally, students in public schools live all nearby their school. In fact, the median for the interquartile range distribution for public schools is 1 kilometer, which means that in half of the public schools the difference in traveled distance between students living close and far from the school is less or equal to 1 kilometer. If we look at subsidized private schools, the median is 1.6 kilometers, and at there are almost no schools where the interquartile range is more than 10 kilometers. In constrast, the median for private school is 3.2 kilometers, which is 2 times the median of subsidized private schools and more than 3 times the median of public schools. Also, there are schools were the interquartile range is 15 kilometers or more, which means that in some schools there are sizable differences in commuting time among students.

Figure 3: Differences in distance to school by school dependence



Source: Own elaboration based in the SIMCE/CIAE data

4 Estimation and results

Now we will briefly explain how we will estimate every part of our theoretical model, and the results of that estimations. To estimate the mode of transport prediction model, we will use the 2012 Origin and Destination survey. Specifically, we will estimate our 7 machine learning models in `Python`, using the `sklearn` machine learning packages, developed by Pedregosa et al. 2011. Our explanatory variables are the student's gender, her household income, the years of schooling of his parents, the traveled distance between the school and the student's house, and the calculated commuting time in the 3 modes of transport. To measure the goodness of fit of the models, we will measure the "accuracy" of every model, that is, the rate of successful predictions over the total of cases.

As were explained in the methodology section, we will measure the accuracy of the prediction of every tuned machine learning model using the 10% of the dataset that was left out. In the table 3 we report these results. It can be seen that the most accurate machine learning model is the random forest decision tree.

Table 3: Accuracy of tuned machine learning models

Model	Accuracy
K-Nearest Neighbors	79.2%
Gaussian Naive Bayes	59.9%
Gradient Boosting Decision Trees	81.2%
Logistic regression	66.5%
Neural Networks	66%
Random Forest	83.75%
Support Vector Machine	80.2%

In the table 4 we report the results of the prediction associated with our best model, the random forest classifier, and we compare it with the original distribution from the Origin and Destination Survey. We can see our prediction model generate a distribution relatively similar to the original distribution. We have to consider that the differences in the distributions of the variables we use in our prediction model can also cause differences in our modes of transport distributions.

Table 4: Predicted mode of transport in the SIMCE dataset

Mode of transport	Original distribution	Random forest prediction
Car	17.2%	15.5 %
Public Transport	30.7%	40.7%
Walking	52.1%	43.8%

With this predictions, we can construct our measures of commuting time, and this will allow us to estimate the equations (1) and (2). To do this, we will use use the average score over the mathematics and spanish test as a measure of academic achievement, and we will use as explanatory variables the student's gender, their parents years of schooling, their household income, school dependence dummies and a full time school dummy.

We will run the model with 2 specifications of the commuting time variable: The predicted commuting time and the traveled distance between the household an the school. For each one of this variables, we will use as the endogenous variable the commuting time to the student's school, and following Kobus et al. 2015 and Tigre et al. 2017, we will use as an instrumental variable the average commuting time to the 2 nearest schools (excluding the student's current school). We will report 4 estimations of each model: In the first column, we will estimate just equation (2) by OLS, in the second column, we will use OLS and school level fixed effects to estimate the same equation. In the third column we will estimate our model (both equations 1 and 2) using instrumental variables, and finally, in the fourth column, we will use our instrumental variables/fixed effects approach to jointly estimate equations (1) and (2).

Regarding the instrumental variables, as was discussed in the Methodology section, our instrument will comply with the relevance condition, because even in the presence of school fixed effects, our instruments influences the school choice decision, thus it affects indirectly the commuting time. Our instrument should also comply with the exclusion restriction, because the commuting time to schools different than the one the student assists should have no influence in the student's academic achievement. Because we are modeling the commuting time using school level fixed effects, we are assuming that there are school level common characteristics. In this context, is logic to think that the error terms too have common components at school level, so we will estimate our errors with school level clusters.

As we have an exactly identified system, we can't test our exclusion restriction, but we can test the relevance condition. Following Baum et al. (2007) and Stock et al. (2002), we will test the relevance condition using the weak instruments test of Stock and Yogo (Stock & Yogo 2005). When testing weak instruments, the first approximation is using the "golden" rule (F>10) from Staiger & Stock (1997), but, as shown in Stock & Yogo (2005), this rule is not very precise when we are dealing with more than one endogenous variable. Because of this, we will use the Stock and Yogo weak instruments test. This test is derived in a setup with homocedastic errors, where the investigator compares Cragg and Donald statistic (Cragg & Donald 1993) with the Stock-Yogo critical values. Since we are estimating a model with cluster in the errors, Baum et al. (2007) suggests that this critical values have to be compared with the Kleibergen and Paap (Kleibergen & Paap 2006) statistic because of its robustness.

The Stock-Yogo test is based in a Wald test with the null hypothesis that the IV estimators are unbiased. The test is based on the asymptotic properties of the rejection rate of the Wald test. We will use the critical value of the test with a 10% maximal IV size, that means, that at worst, the rejection rate of the null hypothesis of unbiased instruments will be 10%. This critical value is 16.38, and will not vary between models.

The table 5 contains the main results of our estimations, where we report the results when we measure commuting time as the traveled distance, and as the predicted commuting time based on the random forest classifier. We can see that in all of our specifications the 4th grade average SIMCE score have a positive and significant effect over the average 8th grade SIMCE score. In almost all of our models the household income and the parents' years of schooling also have a positive and significant effect, and being a woman have a negative and significant effect over the student achievement.

Table 5: Effect of commuting time over average SIMCE scores

Variables	**Traveled Distance**				**Predicted Commuting Time**			
	OLS	**FE**	**IV**	**IV / FE**	**OLS**	**FE**	**IV**	**IV / FE**
Commuting time	0.005***	-0.001	0.066	-0.008	0.001**	-0.000*	-0.000	-0.002
	(0.002)	(0.001)	(0.045)	(0.017)	(0.000)	(0.000)	(0.002)	(0.001)
4th year score	0.686***	0.629***	0.665***	0.628***	0.686***	0.629***	0.688***	0.628***
	(0.008)	(0.007)	(0.017)	(0.007)	(0.008)	(0.007)	(0.010)	(0.006)
Mother years of schooling	0.009***	0.005**	0.004	0.005**	0.009***	0.005**	0.010***	0.005**
	(0.002)	(0.002)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Father years of schooling	0.010***	0.006***	0.004	0.006***	0.010***	0.006***	0.010***	0.005***
	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Household income	0.014***	0.006**	0.008	0.006***	0.015***	0.005**	0.015***	0.004*
	(0.003)	(0.002)	(0.006)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Woman	-0.045***	-0.049***	-0.040**	-0.049***	-0.044***	-0.050***	-0.046***	-0.051***
	(0.017)	(0.011)	(0.017)	(0.010)	(0.017)	(0.011)	(0.017)	(0.011)
Subsidized private	0.053*		0.099**		0.050		0.049	
	(0.031)		(0.045)		(0.032)		(0.033)	
Non-Subsidized private	0.110*		0.194**		0.112*		0.099	
	(0.057)		(0.085)		(0.057)		(0.063)	
Full time school	-0.005		0.053		-0.008		-0.011	
	(0.029)		(0.054)		(0.030)		(0.031)	
Constant	-0.238***	-0.046*	-0.342***		-0.246***	-0.033	-0.219***	
	(0.040)	(0.028)	(0.087)		(0.041)	(0.029)	(0.060)	
Observations	18,215	18,215	18,215	18,215	18,210	18,210	18,210	18,210
Kleibergen-Paap statistic			6.71	39.12			91.42	200.06

If we look at the commuting time effect, we can see that in every OLS specification the commuting time have a positive and significant effect over the academic achievement. When we add fixed effects, we can see that the commuting time variable loss its significance in our traveled distance specification, but with the predicted commuting time, we can see a negative and significant effect. If we recall that the critical value in the Stock-Yogo test of weak instruments is 16.38, we can see that the IV specification for the traveled distance has weak instruments, the IV specification for predicted commuting time has non weak instruments, and both IV-fixed effects specifications have non weak instruments. Anyway, we can see that in both of our IV specifications and our IV/FE specifications the commuting time have a negative effect over academic achievement, but in no specification the effect is statistically significant. Nonetheless, in the IV/FE specification with our predicted commuting time measure, the p-value of our commuting time coefficient is 16.4%, and the coefficient lays between -0.0041 and 0.0007 at a level of confidence of 95%, so even we can't say with certainty that the coefficient is negative, but it is very likely that the effect is negative.

If we interpret this results, we can say that a reduction of 1 minute in commuting time increases the SIMCE average score in 0.002 standard deviations. The within standard deviation of the predicted commuting time is around 19 minutes, so we can think that commuting time can explain differences in achievement of around 0.04 standard deviations within a school.

In the appendix 8.2 we show the first stage results from our main model, and other results. In the table 15 we show and comment our first stages, and in the table 16 we estimate our predicted commuting time model, but separating the students by school dependence. These last models are motivated by figure 3, which shows us that in public and subsidized schools there are no important differences in commuting time among students, but there are important differences in commuting time among students of private schools.

5 Robustness checking and alternative specifications

Now we will do some exercises in order to check if our results are robust. First, we will approach the commuting time problem with a different SIMCE dataset where students self-report their commuting time and their mode of transport usage, but where we don't have their addresses, so we cannnot generate our instrument. The forementioned dataset is the 2004 8th grade SIMCE dataset. In this dataset don't have the same variables as in our main dataset, for instance, we don't have the 4th grade scores for the students, but the commuting time related information can be useful to know the sign of the effect of commuting time or how the students travel.

To start with, in the table 6 we can see the mode of transport distribution in the 2004 8th grade SIMCE dataset. If we compare this distribution with the distributions in the Origin and Destination survey and the 2013 8th grade SIMCE dataset reported in table 4, we can see that the same proportion of the students commute by car in the 3 samples, the proportions of students that walk to school or use public transport are almost identical between the Origin and Destination Survey and the 2004 SIMCE dataset, and are slightly different to the predicted usages in the 2013 SIMCE dataset. This shows us that the dataset used to train our machine learning models is adequate, as the students in that dataset travel similar ways to the students in a real SIMCE dataset.

Table 6: Mode of transport distribution in the 2004 8th grade SIMCE dataset

Mode of transport	**Percentage of students**
Car	18.4%
Public Transport	28.6%
Walking	53%

After this, we can also measure how good is our machine learning prediction by comparing our predicted commuting time with the actual commuting time reported in the 2004 SIMCE dataset. As the data is reported by intervals, we report how many students are in each interval of commuting time in both of our samples in the table 7. We can observe that in both of our samples around the 50% of the students travel less than 15 minutes, around the 80% of the students travel less than 30 minutes, and the other intervals contain similar quantities of students in both samples.

Table 7: Commuting time comparison

Commuting time	**2004 8th grade SIMCE data**	**2013 8th grade SIMCE prediction**
Less than 15 minutes	50.82%	56.01%
Between 16 and 30 minutes	34.80%	20.29%
Between 31 and 45 minutes	8.43%	9.26%
Between 46 and 60 minutes	3.73%	6.55%
Between 61 and 75 minutes	1.16%	3.23%
Between 76 and 90 minutes	0.60%	2.30%
More than 90 minutes	0.46%	2.37%

Finally, we estimate equation (2) with OLS and fixed effects at school level. Here, we cannot use instrumental variables, because without the students addresses we can't calculate the commuting time to the nearby schools. In the table 8 we report the results of both of our estimations, where we use the self-reported commuting time, and we measure achievement as the average standardized SIMCE score. Here we can see that the effect of commuting time appears to be positive and statistically significant when we estimate our model with OLS, but the effect becomes negative when we add fixed effects. These estimations are similar to the ones reported in table 5 when we use our predicted commuting time measure. These results indicate that in Chile we can expect the commuting time effect to be positive when we don't take into account the endogeneity problems, and this appears to be a general fact rather than a particularity of our main dataset.

Table 8: Commuting time comparison

Variables	**OLS**	**FE**
Commuting time	0.004***	-0.001***
	(0.001)	(0.000)
Mother years of schooling	0.048***	0.027***
	(0.002)	(0.001)
Father years of schooling	0.037***	0.019***
	(0.002)	(0.001)
Household income	0.000***	0.000***
	(0.000)	(0.000)
Woman	0.005	0.001
	(0.024)	(0.008)
Subsidized private	0.157***	
	(0.039)	
Non-Subsidized private	0.215***	
	(0.070)	
Constant	-1.209***	-0.521***
	(0.036)	(0.018)
Observations	60,140	60,140
Kleibergen-Paap statistic		

Another way to check the robustness of our results is to show how our model of academic achievement (measured as the average SIMCE score) changes when we change the machine learning model that allows us to build our predicted commuting time measure. In table 9 we report the coefficient, standard deviation and Kleibergen-Paap statistic for our 7 measures of predicted commuting time. The results are reasonably consistent between models: The majority of the OLS models have a positive and significant coefficient, all fixed effects models have a negative coefficient, but only for our K-Nearest Neighbors and Random Forest based measures the effect is significant. If we look at our instrumental variables results, we can see that for every measure the coefficient is not statistically different from zero, and the instrumental variable is not weak. Finally, when we look at our IV/FE specification, we can notice that here all the instruments are not weak, the coefficients of commuting time are all negative, but only the coefficient generated with our K-Nearest Neighbors prediction model is statistically significant.

Table 9: Effect of different measures of predicted commuting time

Measure	**OLS**	**FE**	**IV**	**IV / FE**
K-Nearest Neighbors				
Coefficient	0.000*	-0.000**	-0.000	-0.002**
Standard error	(0.000)	(0.000)	(0.001)	(0.001)
Kleibergen-Paap statistic			81	150
Logistic regression				
Coefficient	0.001***	-0.000	0.000	-0.001
Standard error	(0.000)	(0.000)	(0.002)	(0.001)
Kleibergen-Paap statistic			80	207
Support Vector Machine				
Coefficient	0.001***	-0.000	0.001	-0.000
Standard error	(0.000)	(0.000)	(0.001)	(0.001)
Kleibergen-Paap statistic			94	177
Gaussian Naive Bayes				
Coefficient	0.001**	-0.000	0.002	-0.001
Standard error	(0.000)	(0.000)	(0.002)	(0.001)
Kleibergen-Paap statistic			96	146
Random Forest				
Coefficient	0.001**	-0.000*	-0.000	-0.002
Standard error	(0.000)	(0.000)	(0.002)	(0.001)
Kleibergen-Paap statistic			91	200
Gradient Boosting Decision Trees				
Coefficient	0.001***	-0.000	0.001	-0.001
Standard error	(0.000)	(0.000)	(0.002)	(0.001)
Kleibergen-Paap statistic			159	314
Neural Networks				
Coefficient	0.001***	-0.000	0.002	-0.001
Standard error	(0.000)	(0.000)	(0.002)	(0.001)
Kleibergen-Paap statistic			81	170

There is evidence that suggests that the outcomes of the students are influenced by their school or classmates. Because of this, in our models we control for peer effects or school quality with our school level fixed effects. But if we think of education as a continuous process of human capital accumulation, we may argue that it is not enough to control only with current school fixed effects, as the previous school of students (or more precisely, their previous classmates) may have influenced their current outcomes. To address this issue, we estimate the same specifications that we reported in table 5, but now with a restricted sample of students that have not changed schools between 4th and 8th grade. For these students, we don't need to control for their former schools, because they have not changed schools, so our model can identify with more precision the effect of commuting time over achievement. We report these results in table 10. Here, we can see that for our predicted commuting time measure, the effect of commuting time over academic performance is still negative and now statistically significant in our IV/FE specification, and the effect is almost the double in magnitude that the effect on our main model. We can also notice that now the coefficient of our fixed effects specification is no longer significant.

Table 10: Effect of commuting time over average SIMCE scores, student doesn't change school

Variables	**Traveled Distance**				**Predicted Commuting Time**			
	OLS	**FE**	**IV**	**IV / FE**	**OLS**	**FE**	**IV**	**IV / FE**
Commuting time	0.003	-0.002	0.042	-0.012	0.000	-0.001	-0.000	-0.003*
	(0.002)	(0.002)	(0.029)	(0.019)	(0.000)	(0.000)	(0.003)	(0.002)
4th year score	0.676***	0.659***	0.672***	0.658***	0.676***	0.659***	0.676***	0.658***
	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Mother years of schooling	0.006**	0.005*	0.004	0.005**	0.006**	0.005*	0.006**	0.005*
	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)
Father years of schooling	0.010***	0.007***	0.008***	0.007***	0.010***	0.007***	0.010***	0.006***
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Household income	0.014***	0.006**	0.010**	0.006**	0.014***	0.006**	0.014***	0.004
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Woman	-0.055***	-0.058***	-0.051***	-0.058***	-0.055***	-0.058***	-0.055***	-0.061***
	(0.014)	(0.013)	(0.014)	(0.012)	(0.014)	(0.013)	(0.014)	(0.013)
Subsidized private	0.132***		0.123***		0.132***		0.133***	
	(0.025)		(0.026)		(0.026)		(0.026)	
Non-Subsidized private	0.239***		0.211***		0.241***		0.241***	
	(0.063)		(0.071)		(0.063)		(0.063)	
Full time school	-0.007		-0.001		-0.007		-0.007	
	(0.030)		(0.031)		(0.030)		(0.030)	
Constant	-0.249***	-0.069**	-0.270***		-0.249***	-0.057	-0.246***	
	(0.042)	(0.034)	(0.046)		(0.043)	(0.035)	(0.056)	
Observations	12,237	12,237	12,237	12,237	12,234	12,234	12,234	12,234
Kleibergen-Paap statistic			13.64	53.81			65.01	138.14

Since table 3 shows us that the accuracy of our prediction model is around a 84%, it's possible that the modes of transportation that we predict in our SIMCE dataset are biased. Also, it is possible that even if our model is good, if the sample of students from the origin and destination dataset is really different from our sample of students from the SIMCE dataset, it could be theoretically wrong to use one sample to predict in the other one. Because of these reasons, in our second robustness exercise we will check what happens when we change our predicted modes of transport, and see if with that we can find a significant effect of commuting time over achievement. For this purpose we will run the following Montecarlo experiment:

1. For each student we will draw one random integer numbers between 1 and 3, and this number will be the predicted mode of transport.

2. With this number, we will build our predicted commuting time variable.

3. We can now estimate the commuting time effect with our Fixed Effects-Instrumental Variables approach.

4. We will report if the commuting time variable is significant at a 90% confidence level, and if the instrument pass the weak instrument test.

This experiment will be run 10,000 times. In the table 11 we report the how many of our models have non weak instruments, and if the commuting time coefficient is positive and significant, or negative and significant. We can see that in all of the simulations the instruments were not weak . Regarding the significance of the commuting time variable, we can see that around 5% of the times the variable have a significant effect over the academic achievement, most times the effect is negative, but in the majority of the simulations there was no effect at all. This shows us that, independently of how we estimate the predicted mode of transport for the students, the effect of commuting time is probably negative or zero.

Table 11: Montecarlo experiment results

	Predicted Commuting Time
Non-Weak instruments	100%
Significant and positive	0.25%
Significant and negative	4.85%

Finally, we will check the robustness of our findings in another way: If we look at figure 2, we can see that within every school, the students have very similar commuting times, and maybe because of that lack of variation we don't find a commuting time effect for some specifications. So now we will artificially induce more variation in our data in the following way:

1. Within each school, we will construct the percentiles of the traveled distance's distribution. We will use this measure of commuting time to generate the percentiles because it won't be biased if we get wrong predictions of mode of transport usage.
2. Within each school, we will keep in our sample all the students between the percentiles 1 and 25, and between the percentiles 75 and 100.
3. With this new sample, we will run again all the specifications that we run in the section 4.

The idea of this is to remove the central part of the distance distribution, and with this compare students who are very similar, but have really different commuting times. In the table 12 we can see the results: We now find a negative and significant effect when we use our fixed effects approach, and regarding IV strategies, we only find a negative and significant effect with our IV/FE predict commuting time model. We have to note that in table 5 our commuting time coefficient for this model was -0.002, and now is -0.003, a result that is consistent with the logic of the robustness exercise: Because now we compare students who live really close to their school with students who live really far, the commuting time now have a more negative effect.

6 Magnitude of the effect and comparison

Now we are going to discuss the magnitude of the previously found effect of commuting time over achievement, and then we will compare the magnitude of our effect with the effects reported in other articles regarding commuting time, and with the results of other articles regarding other topics in education.

First, in our main specification with instrumental variables and fixed effects reported in table 5, the coefficient of commuting time is -0.002, which means that an increase of 1 minute in the expected commuting time reduces the average SIMCE score by 0.002 standard deviations. To make our results comparable with other articles, we can say that increasing the commuting time in one *within-school* standard deviation (20 minutes) reduces the average SIMCE score by around 4% of a standard deviation. As our commuting time coefficient is not statistically significant, our effect could be zero. We report the effect of increasing a within-school standard deviation of commuting time rather than the effect of the overall

Table 12: Robustness check: Effect of commuting time over average SIMCE scores

Variables	**Traveled Distance**				**Predicted Commuting Time**			
	OLS	**FE**	**IV**	**IV / FE**	**OLS**	**FE**	**IV**	**IV / FE**
Commuting time	0.002	-0.001	0.025	-0.017	0.000	-0.000	-0.002	-0.003**
	(0.002)	(0.002)	(0.025)	(0.015)	(0.000)	(0.000)	(0.002)	(0.002)
4th year score	0.680***	0.619***	0.674***	0.618***	0.680***	0.619***	0.684***	0.618***
	(0.010)	(0.010)	(0.012)	(0.010)	(0.010)	(0.010)	(0.011)	(0.010)
Mother years of schooling	0.011***	0.007**	0.009***	0.007**	0.011***	0.007**	0.011***	0.006**
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Father years of schooling	0.008**	0.001	0.005	0.002	0.008***	0.001	0.008***	-0.000
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Household income	0.018***	0.012***	0.015***	0.012***	0.018***	0.012***	0.018***	0.009***
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Woman	-0.046**	-0.058***	-0.043**	-0.059***	-0.046**	-0.058***	-0.050**	-0.063***
	(0.021)	(0.017)	(0.021)	(0.016)	(0.021)	(0.017)	(0.021)	(0.016)
Subsidized private	0.074**		0.089**		0.073**		0.069*	
	(0.035)		(0.038)		(0.035)		(0.037)	
Non-Subsidized private	0.088		0.116*		0.089		0.071	
	(0.065)		(0.069)		(0.065)		(0.069)	
Full time school	-0.014		0.005		-0.015		-0.019	
	(0.033)		(0.038)		(0.033)		(0.035)	
Constant	-0.246***	-0.038	-0.297***		-0.252***	-0.028	-0.193**	
	(0.049)	(0.043)	(0.074)		(0.050)	(0.044)	(0.080)	
Observations	8,427	8,427	8,427	8,293	8,424	8,424	8,424	8,290
Kleibergen-Paap statistic			13.69	36.21			75.43	102.90

standard deviation, because with our fixed effects we are comparing students within the same school, so the effect of a within-school standard deviation of commuting time can tell us how much of the differences in performance within the 8th graders of a school could be explained by commuting time.

In one of our alternative specifications reported in table 10, we keep in our sample only the students are in the same school in the 4th and 8th grade. In that specification the commuting time coefficient is -0.003 and is statistically significant. For this sample of students, the within-school standard deviation is 14.5 minutes, so an increase in commuting time of 14.5 minutes could explain a reduction in academic performance of 4.35% of a standard deviation. Also, in another specification, when we split our sample by school dependence, as reported in table 16, the commuting time coefficient of our private school students is -0.01. Here, the within-school standard deviation for students of private schools is 13.5 minutes, so an increase in one within-school standard deviation of commuting time could cause a reduction of 13% of a standard deviation in academic achievement. But again, this coefficient is not statistically significant, so our effect could be zero. Summarizing, increasing the commuting time in one within-school standard deviation could reduce the performance of a student by between 0% and 13% of a standard deviation.

Now that we have explained the magnitude of our commuting time effect, we have to compare it to other related articles. The most related paper is Tigre et al. (2017). They found, with different causal instrumental variables specifications, that an increase of 1 minute of commuting time causes a reduction of between -0.012 and -0.005 standard deviations in the test scores for 6th graders in public schools in Brazil. In the article, the standard deviation of commuting time is 19 minutes, and that implies that increasing the commuting time in one standard deviation could diminish the test score by between 10% and 24% of a standard deviation. Another related paper is Kobus et al. (2015), where the authors claim, using data from university students in The Netherlands, that a standard deviation increase in commuting time reduces the average grade by about 33% of a standard deviation.

Finally, we can compare the magnitude of our effect with the results found in other relevant papers in education. For instance, Angrist and Lavy (1999) studies the effect of classroom size over achievement. They state that reducing the classroom size by 1 standard deviation could increase the performance of the students by at least 13%. Moreover, the investigators compare their results to others articles regarding classroom size, and they claim that other articles, like Finn and Achilles (1990), found even bigger classroom size effects. Another interesting paper is Duflo and Hanna (2005), where the effect of teacher absenteeism over performance is addressed. They found that a program to reduce teacher absenteeism increased the performance of the students by 17% of a standard deviation.

In conclusion, our commuting time effect can explain between 4% and 13% of a standard deviation in achievement, whereas other articles found commuting time effects of up to 24% or 33% of a standard deviation, and other papers in education found effects of more than 13%. So even if our effect were statistically significant, it would not be the most important effect documented in education, but it would still be important.

7 Concluding remarks

We study the effect of commuting time over academic achievement with a sample of 8th graders from Santiago. We estimate an auxiliary mode of transport machine learning model to predict the mode of transport usage in our sample, and with these predictions we generate proxies for the commuting time of our individuals. We estimate the effect of commuting time over scholar performance with an instrumental variables model (Kobus et al. 2015; Tigre et al. 2017), and we also add fixed effects at school level to control for school choice.

We find that, when we don't account for the endogeneity of the commuting time variable, specifically for our school selection problem, the effect of the commuting time over achievement appears to be positive, a result that contradicts all the previous international literature, and that is valid for different SIMCE datasets. But when we apply our instrumental variables and fixed effects approach, we find that commuting time have a negative effect over the academic achievement of the students, but this effect is not always statistically significant. In order to check the robustness of our results, we have changed our prediction model, generated random transportation modes with a Monte Carlo experiment, modified the distance to school distribution within every school, and constrained our sample to students that study in the same school in 4th and 8th grade, and in general the evidence suggests that the effect of commuting time over academic achievement is negative or zero, but never positive.

Regarding the magnitude of the commuting time effect, looking at our different specifications and results for different subsamples, we can say that increasing the commuting time in one within-school standard deviation could diminish the average standardized SIMCE score of the students by between 4% and 13% of a standard deviation. This effect is relevant, but as our coefficients are not always significant, the effect could be zero. Moreover, even if the effect were statistically significant, other articles regarding commuting time or other variables related to educational processes generally found effects bigger than the one we found.

This article contribution is not only limited to the estimation of the commuting time effect, but it also makes a methodological contribution to the commuting time literature: First, we combine our machine learning approach with our commuting time estimations made with Google Maps to generate a proxy of commuting time that is much more accurate than just using the linear distance. Another contribution is adding the school level fixed effects to the typical instrumental variables model used in the literature, as this mitigates the school choice problem in an educational system as complex as the Chilean.

Despite all of these contributions, our article still have room for improvement: For instance, with data that shows the mode of transport the students use, our mode of transportation prediction model could be eliminated, and the effect of commuting time over achievement could be estimated more precisely. Also, perhaps a more complex methodology involving structural equations could model the school choice process, thus eliminating the necessity of using fixed effects. This approach could also unify the literature about commuting time and school choice with the articles relating commuting time and academic performance.

References

- Asahi, K. (2016). Closer Proximity to the Subway Network Implies Lower High School Test Scores: Evidence from a Subway Expansion.

- Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. The Quarterly Journal of Economics, 114(2), 533-575.

- Athey, S. (2017). The impact of machine learning on economics. In Economics of Artificial Intelligence. University of Chicago Press.

- Banerjee, A., & Duflo, E. (2006). Addressing absence. Journal of Economic perspectives, 20(1), 117-132.

- Baum, C. F., Schaffer, M. E., & Stillman, S. (2007). Enhanced routines for instrumental variables/GMM estimation and testing. Stata Journal, 7(4), 465-506.

- Becker, S. O., & Woessmann, L. (2009). Was Weber wrong? A human capital theory of Protestant economic history. The Quarterly Journal of Economics, 124(2), 531-596.

- Chumacero, R. A., Gómez, D., & Paredes, R. D. (2011). I would walk 500 miles (if it paid): Vouchers and school choice in Chile. Economics of Education Review, 30(5), 1103-1114.

- Ciccone, A., & Papaioannou, E. (2009). Human capital, the structure of production, and growth. The Review of Economics and Statistics, 91(1), 66-82.

- Colclough, C. (1982). The impact of primary schooling on economic development: a review of the evidence. World Development, 10(3), 167-185.

- Cragg, J. G., & Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. Econometric Theory, 9(02), 222-240.

- Currie, J., & Moretti, E. (2003). Mother's education and the intergenerational transmission of human capital: Evidence from college openings. The Quarterly Journal of Economics, 118(4), 1495-1532.

- Dee, T. S. (2004). Are there civic returns to education?. Journal of Public Economics, 88(9), 1697-1720.

- Dee, T. S. (2007). Teachers and the gender gaps in student achievement. Journal of Human Resources, 42(3), 528-554.

- Dickerson, A., & McIntosh, S. (2013). The Impact of Distance to Nearest Education Institution on the Post-compulsory Education Participation Decision. Urban Studies, 50(4), 742-758.

- Duflo, E., & Hanna, R. (2005). Monitoring works: Getting teachers to come to school (No. w11880). National Bureau of Economic Research.

- Falch, T., Lujala, P., & Strøm, B. (2013). Geographical constraints and educational attainment. Regional Science and Urban Economics, 43(1), 164-176.

- Ferreyra, M. M. (2007). Estimating the effects of private school vouchers in multidistrict economies. American Economic Review, 97(3), 789-817.

- Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. American Educational Research Journal, 27(3), 557-577.

- Fryer Jr, R. G., & Levitt, S. D. (2010). An empirical analysis of the gender gap in mathematics. American Economic Journal: Applied Economics, 2(2), 210-40.

- Gallego, F. A., & Hernando, A. (2009). School choice in Chile: Looking at the demand side.

- Gibbons, S., & Vignoles, A. (2012). Geography, choice and participation in higher education in England. Regional science and urban economics, 42(1-2), 98-113.

- Hanushek, E. A., & Wößmann, L. (2007). The role of education quality for economic growth. NJ: Princeton University Press

- Kjellström, C., & Regnér, H. (1999). The effects of geographical distance on the decision to enrol in university education. Scandinavian Journal of Educational Research, 43(4), 335-348.

- Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. Journal of econometrics, 133(1), 97-126.

- Kobus, M. B., Van Ommeren, J. N., & Rietveld, P. (2015). Student commute time, university presence and academic achievement. Regional Science and Urban Economics, 52, 129-140.

- Lochner, L., & Moretti, E. (2004). The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. American Economic Review, 155-189.

- Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. The Economic Journal, 121(552), 463-484.

- Méndez, M.L. & Gayo, M. (2018) Upper Middle Class Social Reproduction: Wealth, Schooling, and Residential Choice in Chile, Palgrave Pivot Series, New York.

- Milligan, K., Moretti, E., & Oreopoulos, P. (2004). Does education improve citizenship? Evidence from the United States and the United Kingdom. Journal of public Economics, 88(9), 1667-1695.

- Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. Journal of Economic Perspectives, 31(2), 87-106.

- Paredes, V. (2014). A teacher like me or a student like me? Role model versus teacher bias effect. Economics of Education Review, 39, 38-49.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. Econometrica, 73(2), 417-458.

- Sá, C., Florax, R. J., & Rietveld, P. (2006). Does accessibility to higher education matter? Choice behaviour of high school graduates in the Netherlands. Spatial Economic Analysis, 1(2), 155-174. and segregation by race and poverty. Social Problems, Vol. 50(2); pp. 181-203.

- Staiger, D. O., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. Econometrica, 65(3), 557-586.

- Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics, 20(4), 518-529.

- Stock, J. H., & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, 80.

- Temple, J. (2002). Growth effects of education and social capital in the OECD countries. Historical Social Research/Historische Sozialforschung, 5-46.

- Tigre, R., Sampaio, B., & Menezes, T. (2017). The Impact of Commuting Time on Youth's School Performance. Journal of Regional Science, 57(1), 28-47.

- Westman, J., Olsson, L. E., Gärling, T., & Friman, M. (2015). Children's travel to school: satisfaction, current mood, and cognitive performance. Transportation, 1-18.

8 Appendix

8.1 Data Loss

In this appendix we analyze why our sample is so little, considering it is a sample of all the 8th grade students from the "Gran Santiago". In the table 13 we can see how many observations we lose in each step when we build our dataset: We start with 255,132 observations. If we clean the database (drop observations without gender, duplicates, no score in the test), we keep 203,791 students. Then, if we add the students and parents questionnaires (that contain the parents' years of schooling, household income and other relevant variables) our sample is reduced to 164,290 observations. When we consider only the students living in the Metropolitan Region (where Santiago is located) we keep only 61,162 observations.

When we add the commuting time and distance data provided by the CIAE, our sample is reduced to 24,168 students, and then our sample is reduced to 22,858 when we add our machine learning predictions. When we later restrict the sample to the students in the "Gran Santiago"[9] we maintain the same sample of 22,258, and finally, when we include in the dataset the 4th grade SIMCE scores, the sample is reduced to 18,342 observations.

Table 13: Data loss

Dataset	**Observations**
Full SIMCE 2013 dataset	255,132
Valid observations	203,791
Valid parents' responses	164,910
Valid student's responses	164,290
Lives in the Metropolitan Region	61,162
Has commuting time data	24,168
Has predicted mode of transport	22,858
Lives in the Gran Santiago	22,858
Have 4th grade scores	18,342

Because we lost more than half of our sample when we added the commuting time and machine learning data, now we are going to analyze how representative is our merged data. For this purpose, in the table 14 we make a comparison of the mean of several variables in our different samples. We can see that the

[9] We can't do this earlier, because the student address is not available in the original SIMCE dataset.

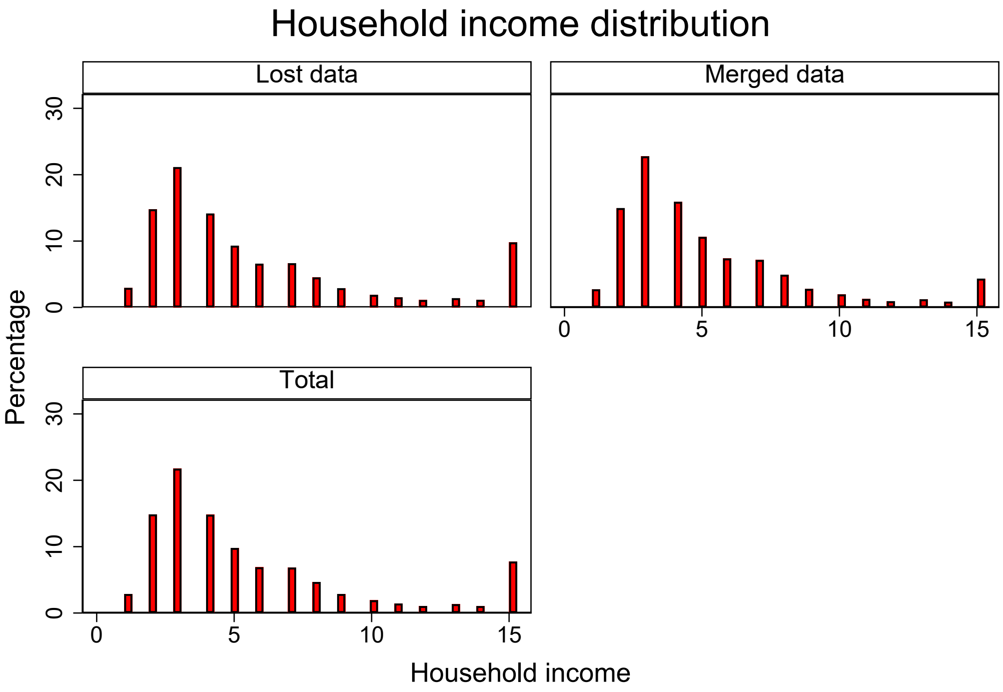
test scores are really similar (all of them are standardized), the gender composition is the same in the 3 samples, and the parents years of schooling is very similar.

Table 14: Sample comparison

Variable	**Merged sample**	**Lost sample**	**Full sample**
Math. Score	-0.03	0.02	0.00
Spanish score	-0.03	0.02	0.00
Average score	-0.03	0.02	0.00
Woman	0.50	0.50	0.50
Mother years of schooling	11.73	11.99	11.89
Father years of schooling	11.80	12.07	11.97
Observations	**22,858**	**38,304**	**61,162**

We can see that the means of all the included variables are similar across samples. Next, in the figure 4 we compare the household income distribution across our samples. We can see that the distributions are similar, but the merged data have less people in the last income group. With this, we can say that our merged sample is relatively representative of the SIMCE dataset for the Metropolitan Region.

Figure 4: Household income comparison



Source: Own elaboration based in the SIMCE/CIAE data

8.2 First stages and other results

First, we will report the first stages of our instrumental variables and our instrumental variables/fixed effects specifications. The table 15 contains the first stages of the models reported in table 5. We can notice that our instrument, the average commuting time to the 2 nearest schools, always has a positive and significant impact over our endogenous variable, the commuting time to the actual school, a finding that is consistent with the fact that our instrument generally is non weak. If we look at the rest of the coefficients, we can see that if we compare the IV specifications of both of our commuting time measures, the signs of the variables are generally similar. Despite that, when we add fixed effects in both of our models, the signs and significance level of our variables change dramatically, because the fixed effects in the main equation explain some of the endogeneity associated with the commuting time, but also because now we have also added fixed effects in our first stage. Indeed, now in the first stage we also control for unobserved characteristics that led parents with children in the same school to choose that school, and not to choose the nearby schools.

Summarizing, here it is not really important to look at the signs of the coefficients of our first stage, but it is important to see that they change because our identification strategy with fixed effects change the role of the instrumental variables in the classic model estimated in the literature.

Now, in the table 16 we report the results of our main model, but separating the sample by school dependence. Here we have some interesting results: First, we can see that the commuting time effect estimated by OLS is only positive and significant for the students in public schools. This could mean that, for that subsample of the population, we find a positive effect when we don't account for the endogeneity, and could be interpreted as evidence that going to a far away public school pays, probably because the local supply of schools that the household faces it is not very good.

Another interesting result is that if we look at our IV/FE commuting time coefficients, we can see that they differ greatly in magnitude. The effect for subsidized private schools is so little that when we look at 3 decimal places we still see only zeros, the coefficient for public schools is -0.003, which is similar to the -0.002 coefficient of our regular model, but the coefficient for private school is -0.01, with is more

Table 15: First stages of instrumental variables models

Variables	Traveled Distance		Predicted Commuting Time	
	IV	**IV / FE**	**IV**	**IV / FE**
Average commuting time	0.420***	0.915***	0.662***	0.933***
	(0.162)	(0.146)	(0.069)	(0.066)
4th year score	0.334***	-0.061*	1.937***	-0.271
	(0.087)	(0.033)	(0.418)	(0.184)
Mother years of schooling	0.087***	0.002	0.429***	-0.182***
	(0.015)	(0.012)	(0.072)	(0.067)
Father years of schooling	0.089***	0.015	0.148**	-0.343***
	(0.012)	(0.011)	(0.069)	(0.065)
Household income	0.107***	0.001	-0.318***	-0.775***
	(0.020)	(0.015)	(0.086)	(0.083)
Woman	-0.082	-0.002	-1.385*	-1.054***
	(0.171)	(0.060)	(0.774)	(0.318)
Subsidized private	-0.742**		-1.622	
	(0.295)		(1.341)	
Non-Subsidized private	-1.512*		-11.136***	
	(0.874)		(3.174)	
Full time school	-0.963**		-3.056*	
	(0.424)		(1.800)	
Constant	1.425***		16.414***	
	(0.493)		(2.314)	

than 5 times the coefficient of our general model. So, this could be evidence that the effect of commuting time is greater for students in private schools, and this make sense, because these students are the most heterogeneous in terms of commuting time.

Sadly, we can't be sure about our results for the private schools sample, because we only have 1,084 observations, and we loss 166 degrees of freedom when we add our fixed effects by school. Furthermore, our instrumental variable is weak in both our IV and IV/FE private schools specifications[10], so our coefficients might be biased. This makes our coefficient not very reliable, and we can't conclude anything about this results.

Table 16: Effect of commuting time over achievement by school dependence

Variables	**Public schools**				**Subsidized private schools**				**Private schools**			
	OLS	**FE**	**IV**	**IV / FE**	**OLS**	**FE**	**IV**	**IV / FE**	**OLS**	**FE**	**IV**	**IV / FE**
Commuting time	0.002***	-0.000	0.000	-0.003	-0.000	-0.000	0.000	-0.000	-0.001	-0.003**	-0.004	-0.010
	(0.000)	(0.000)	(0.004)	(0.002)	(0.000)	(0.000)	(0.002)	(0.001)	(0.001)	(0.001)	(0.010)	(0.008)
4th year score	0.686***	0.609***	0.694***	0.608***	0.674***	0.637***	0.674***	0.637***	0.707***	0.666***	0.708***	0.670***
	(0.017)	(0.011)	(0.023)	(0.011)	(0.009)	(0.008)	(0.009)	(0.008)	(0.031)	(0.031)	(0.031)	(0.029)
Mother years of schooling	0.016***	0.005	0.017***	0.005	0.004	0.004	0.004	0.004	0.008	0.012	0.007	0.008
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.010)	(0.010)	(0.013)	(0.012)
Father years of schooling	0.011***	0.007**	0.012***	0.006**	0.008***	0.005**	0.008***	0.005**	0.015	0.007	0.012	-0.001
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.009)	(0.010)	(0.012)	(0.010)
Household income	0.025***	0.007*	0.025***	0.004	0.013***	0.005	0.014***	0.005	0.012	0.004	0.008	-0.005
	(0.005)	(0.004)	(0.005)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.008)	(0.007)	(0.017)	(0.011)
Woman	-0.062	-0.078***	-0.065*	-0.081***	-0.039**	-0.039***	-0.038**	-0.039***	-0.008	-0.034	-0.018	-0.057
	(0.039)	(0.019)	(0.039)	(0.019)	(0.015)	(0.014)	(0.016)	(0.013)	(0.038)	(0.040)	(0.053)	(0.049)
Full time school	0.000		-0.020		0.019		0.018					
	(0.053)		(0.063)		(0.034)		(0.035)					
Constant	-0.403***	-0.126***	-0.350***		-0.097*	-0.000	-0.108		-0.152	0.102	0.010	
	(0.069)	(0.043)	(0.102)		(0.051)	(0.038)	(0.067)		(0.220)	(0.176)	(0.683)	
Observations	5,828	5,828	5,828	5,828	11,298	11,298	11,298	11,298	1,084	1,084	1,084	1,084
Kleibergen-Paap statistic			20.06	58.27			76.94	148.56			4.81	8.19

[10] Perhaps a good idea would be to calculate a new instrumental variable defined as the average commuting time to the 2 nearest private schools, but these data is not currently available to us.