

Educational choice and inter-regional migration – The causal effect of high school on moving out of non-urban areas¹

Elise Stenholt Sørensen² and Anders Holm³

Abstract:

This paper identifies a major source of migration from rural to urban areas in Denmark. We estimate the causal effect of obtaining a high school degree on the decision to leave rural areas and move to urban areas. The net-migration from rural to large urban areas has increased significantly, leaving rural areas behind with a declining population and problems with retaining human capital. The main driver behind this is an increasing share of young adults migrating towards large urban areas. Because young people who migrate from rural areas to larger cities typically do not return, the location decisions of young people have long-run implications for regional inequality regarding human capital and regional economic growth. This paper answers the question on the causal effect of completing high school on out-migration from rural areas. The study is based on panel data from Danish administrative registers. When employing an IV-approach, we find that completing high school increases young adults' probability of moving out of a rural area with 65 percentage point. Furthermore, we find that the causal effect of completing high school is heterogeneous across socioeconomic background. The causal effect of completing high school is largest for young adults with low socioeconomic background.

JEL classification: I20, J24, J61, R23, R58

Keywords: interregional migration, upper secondary education, human capital, graduate migration, regional development, distance to school

¹ The authors thank Bo Honoré, Robert Andersen, Alessandra Faggian, Thomas Crossley for reading previous versions of the paper and giving their helpful suggestions and comments. We also thank Steven Durlauf, Don Davis, Nikolai Kuminoff, Rasmus Landersø, Ismir Mulalic, Lars Pico Geerdsen, Lars Winther and Lars Nesheim for their constructive comments.

² Kraks Fond – Institute for Urban Economic Research and Department of Geosciences and Natural Resource Management, University of Copenhagen

³ Department of Sociology, University of Western Ontario, Department of Economics, University of Western Ontario and Kraks Fond – Institute for Urban Economic Research

1. Introduction

This paper identifies a major source of migration from rural to urban areas in Denmark. We are the first to estimate the causal effect of obtaining a high school degree⁴ on the decision to leave rural areas and move to urban areas.

Our study aims at estimating the causal effect of secondary education on moving out of non-urban areas, using administrative data from Statistics Denmark in combination with detailed measures of distance to high school institutions. The data has rich information about five birth cohorts⁵ of Danish 15-years old, who grew up in a rural area. In order to take endogeneity issues into account, regarding the choice of high school, we use several identification strategies. First, we use distance to high school as an instrumental variable. As parental location in relation to access to high school may not be exogenous, we also employ a second IV strategy where we use the fact that some high school opens or closes after the parents has located into their area of residence. We find very large local average treatment effects (LATE) of high school on moving out of rural areas. Furthermore, we find that the causal effect of completing high school is heterogeneous, so the causal effect is largest for young adults with low socioeconomic background. To get an idea of the size of the average treatment effect we also employ a sibling and twin fixed effect analysis where we use changes in choice of education among siblings.

Since the industrial revolution, there have been major changes in the geographical distribution of population across Europe. Urban areas are now home to almost three quarters of the European Unions (EU's) population (Eurostat, European Union, 2016). The same picture is seen in the US where about 80 percent of the population live in an urban or suburban area (United States Census Bureau, 2016) . Urbanisation has become a global phenomenon, accounting for an increasing share of economic growth. Nowadays, the process is particularly evident in emerging economies and the developing world (United Nations, 2018). The net-migration from rural to large urban areas has increased significantly during the last decades, also in developed countries (United Nations, 2015; Van Der Gaag & Van Wissen, 2008). The increased migration is leaving rural areas behind with a

⁴ In this paper we use the term high school to denote the Danish equivalent of “gymnasium”. In Denmark, high school (or gymnasium) is an upper secondary academic educational option after compulsory school. Student are in the age group 15-16 years when they decide to enter high school in Denmark. Around half a cohort chooses high school while the reminder either enter a vocational education (physically different school than gymnasium) or drops out of the educational system.

⁵ We can only use the five most recent cohorts because compulsory school grades are not available for earlier cohorts.

declining population and problems with retaining human capital. This causes increased regional divergence in economic progress and prosperity.

The inequality between regions in the European Union (EU) and in the US has increased significantly during the last decades (Iammarino, Rodriguez-Pose, & Storper, 2018; Moretti, 2013). Glaeser (2013) argues that there has been remarkable geographical heterogeneity in the past, but the big change today is the regional divergence in human capital. In the US a growing share of residents ages 25 and older have graduated from college, but the growth has been much larger in the urban areas: In 2016, 35% of the urban residents had a bachelor's degree or more education, compared with 19% in rural areas. In 2000 the numbers were 28% and 15% in urban and rural areas, respectively (Pew Research Center, 2018). The same picture is seen in the EU where 37% of the population aged 25-64 living in cities had a tertiary education, compared with only 21% of those living in rural areas (Eurostat, European Union, 2016).

In Denmark, we find that a main driver behind the increased regional inequality between rural and urban areas is an increasing share of young adults migrating towards large urban areas, which we will elaborate in section 2. Because young people who migrate from rural regions to larger towns typically do not return, the location decisions of young people have long-run implications for the population and the age structure in both less-urban areas (Berck, Tano, & Westerlund, 2016, p. 2) as well as urban areas. Therefore, youth migration has implications for regional inequality regarding human capital, which again affects the possibilities of regional economic growth.

If those migrating have more valuable skills than those staying behind the net out migration may leave the local labor market deprived of valuable human capital. If human capital has synergy effects (Moretti, 2013; Rosenthal & Strange, 2008) it may even leave those on the local labour market with less value of their own human capital.

This leaves the natural question of, whether the increasing economic inequality among regions in Europe and the US is a problem? Iammarino, Rodriguez-Pose, and Storper, 2018 argue that the *“regional economic divergence has become a threat to economic progress, social cohesion and political stability in Europe”* and is one of the reasons to why the political landscape has become so divided between rural and urban populations in the Western World. On the other hand, cities are motors of the EU economy, providing hubs for wealth creation and agglomeration economies (Eurostat, European Union, 2016).

However, we offer one potential important driver of the rural urban divide. We find very large effects on net out migration of youth with high school education compared to youth with no

education beyond compulsory education or youth with vocational education. Since our analysis is causal, we point to a mechanism. If the adolescent had not achieved a high school education, then they would not have migrated.

We use administrative data from Statistics Denmark in combination with detailed measures of the distance to high school institutions. The data has rich information about a large range of birth cohorts of Danish 15-years old, who grew up in a rural area. We follow each individual over at 10-year time period and analyze the timing of their first move away from home. Furthermore, we use the distance between the place of residence to the nearest high school institution as instruments for choice of high school, as choice of upper secondary education could be endogenous. We show that obtaining high school degree increase outmigration rates of rural areas by 65 percentage point. As parental location in relation to access to high school may not be exogenous, we also employ a second IV strategy where we use the fact that some high schools opens or closes down. Observing that adolescent youth is the only age bracket in Denmark with net migration out of rural areas, we identify a major factor driving rural urban migration.

We further find that our instrument (distance to nearest high school institution) has a larger effect on students from families with lower socio-economic status than on students with a stronger parental background. This makes sense since students with lower educated parents are presumably more on the margin of entering high school.

The remainder of the paper is organized as follows. Section 2 gives a descriptive overview of migration patterns and educational attainment of adolescent youth in Denmark. Section 3 reviews the literature on adolescent migration, while section 4 offers an overview over the Danish educational system relevant for our analysis. In section 5 we describe the data used in our empirical analysis, and section 6 presents a theoretical model for an adolescent's decision to enter high school and to migrate out of a rural area and into an urban area. This section motivates our empirical findings. In section 7 we discuss our empirical framework and section 8 presents the results of our analysis. Finally, section 9 gives some concluding remarks.

2. Descriptive overview

The net-migration from rural to large urban areas has increased significantly during the last decades, also in developed countries (United Nations, 2015; Van Der Gaag & Van Wissen, 2008).

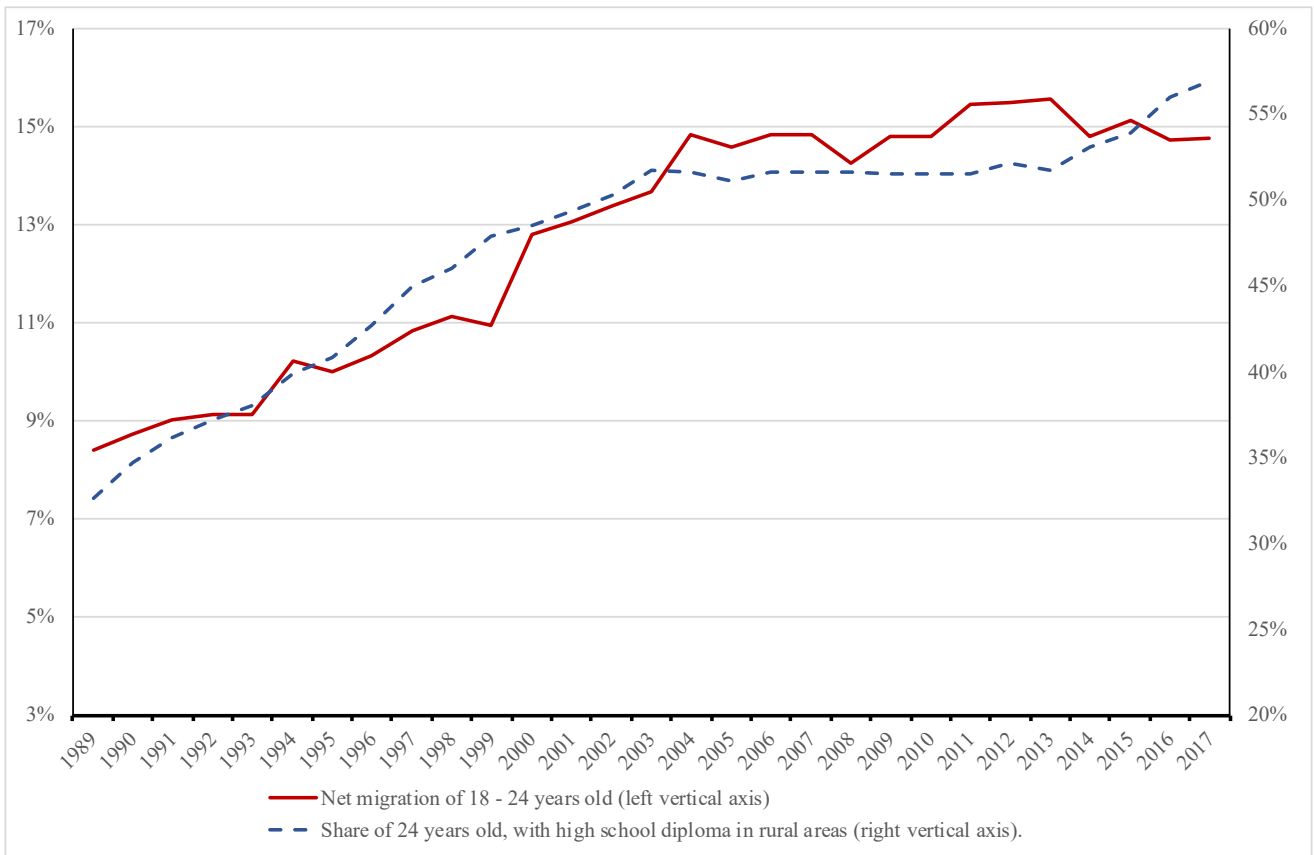
Understanding the reasons behind this may be interesting in order to determine the welfare change for both those leaving and those staying behind. Locally, the effects of net-migration may depend on the

skill composition of those staying behind and those migrating. If those migrating have more valuable skills than those staying behind the locally net out migration may leave the local labor market deprived of valuable human capital. If human capital has synergy effects (Rosenthal & Strange, 2008) migration may even leave the those on the local labour market with less value of their own human capital. Looking at the Danish case, we see that the main driver behind out-migration from rural areas is an increasing share of young adults migrating towards the large urban areas. Simultaneously, the share of young adults completing a high school have increased significantly, both in rural and urban areas. Figure 1 below, illustrates how the growth in youth migration towards urban areas largely follow the same trend as the growth in educational attainment, from 1989 to 2017. The red solid line in Figure 1 shows the share of young adults, 18-24 years old, who moves to a large city, out of the total share of young adults living in a rural area. A rural area is defined as cities with less than 20.000 inhabitants and rural areas.⁶ We find that net outmigration from rural areas among youth has increased from 8 percent in 1989 to 15 percent in 2017⁷. At the same time, the share of young adults, aged 24 years, completing high schools, have increased from 33 percent in 1989 to 57 percent in 2017 (the blue dashed line in Figure 1).

Figure 1: Growth in net out-migration from rural areas and educational attainment among youth, 1989 – 2017

⁶ We make robustness checks of our definition of the dependent variable. These checks include defining a move as a move to the four largest cities in Denmark and only moves to cities with a university. We return to this.

⁷ In absolute numbers, the outmigration from rural areas, among 18-24 years old, have increased from 25,918 individuals in 1989 to 32.587 individuals in 2017. At the same time, the total number of youth living in rural areas, has decreased . It has declined from 230,015 individuals in 1989 to 181,559 young individuals in 2017.



Sources: Administrative registers from Statistics Denmark. Net migration is calculated using a full population sample of 18-24 years old and information of moving decisions each year. The share with high school diploma is calculated among the full cohort of 24 years old each year. The population is divided in rural and urban population on the basis of their residential address, when they were 17 years old. Persons without residential address in Denmark when they were 17 and 24 is excluded from the analysis.

Therefore, is it conceivable that growth in educational attainment among young adults is one of the main drivers behind the increasing youth migration towards the larger cities. From a regional policy perspective, it is highly relevant to reveal, if the correlation between education choice and migration among youth, is in fact a causal relationship. If so, an unintended consequence of investing in increased access to high schools in rural areas, could be that young adults move away from such areas, after graduation.

Before proceeding it may be natural to ask if the outmigration shown in Figure 1 above is not replaced by an in-migration of similar size such that the share of people remains unchanged in rural and non-urban areas. In Figure 2 below we show the net migration rate out of rural areas, by age groups. The net migration rate is the difference between the number of immigrants (people coming

into an area) and the number of emigrants (people leaving an area) throughout the year. A negative net migration rates indicates that there are more people leaving an area than entering an area.

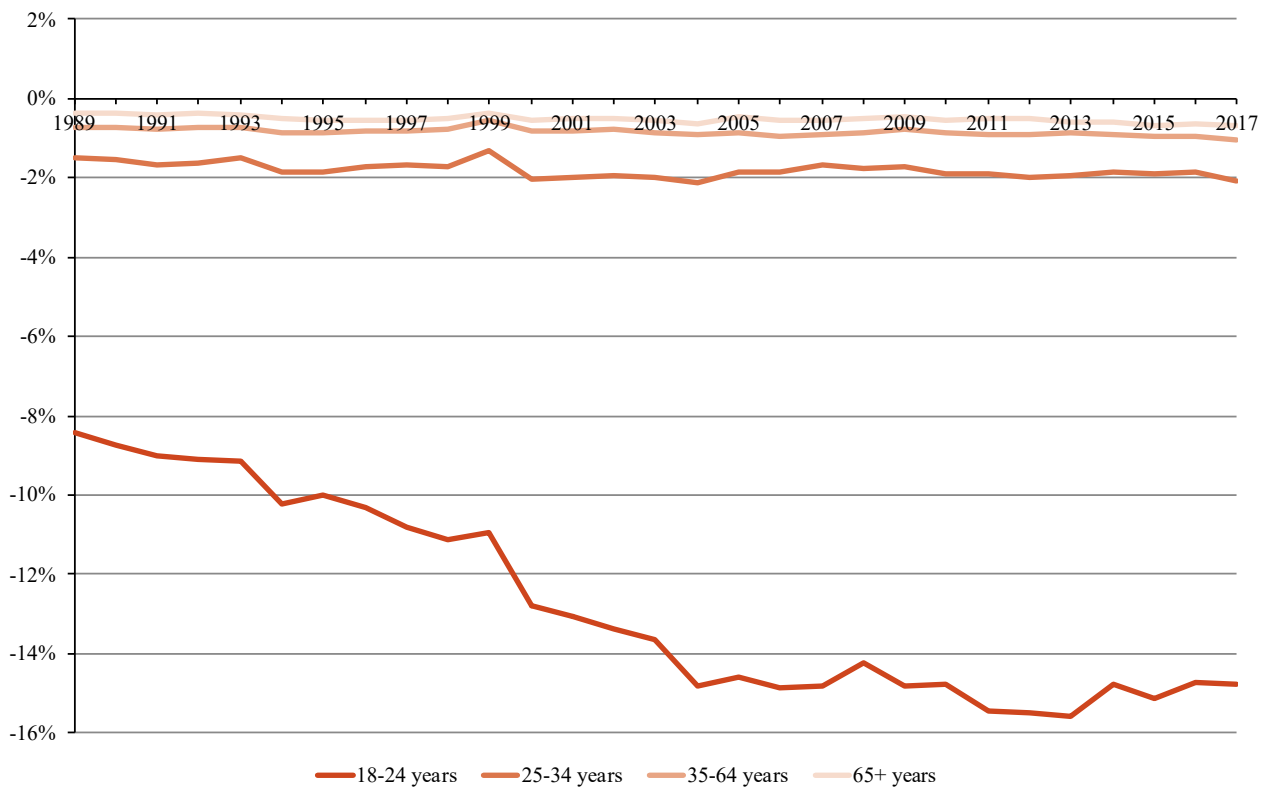
Net migration rate in rural areas is calculated using the formula below, for each age group:

$$N = \frac{I - E}{M} * 100$$

Where N denotes Net migration rate in percentage points. I denotes the number of immigrants (young people moving in) and E denotes the number of emigrants (young people between 18 – 24 years moving out). M denotes mid year population of young people (18 – 24 years), that is:

$$M = \frac{\text{Population at start of year} + \text{Population at end of year}}{2}$$

Figure 2. Net migration rates in rural areas, by age group, 1989 – 2017



Sources: Administrative registers from Statistics Denmark. Net migration is calculated using a full population sample of people living in rural areas and information of moving decisions each year.

Figure 2 illustrates net-migration rates by age groups. From the figure it is evident that net-outmigration from rural areas is driven by the age group 18-24 years old.⁸ We see that the increasing net out-migration from rural areas is almost solely driven by the age group 18-24 years old. For older age groups, net migration from rural areas is also negative, but much smaller and constant over time. As briefly discussed in the introduction, because young people who leave rural regions in favour of larger cities typically do not return, as illustrated above, the location decisions of young people have long-run implications for the population and the age structure in both less-urban areas as well as urban areas (Berck et al., 2016, p. 2). Therefore, youth migration has implications for regional inequality regarding human capital, which again affects the potential of regional economic growth.

3. Previous empirical evidence

The previous empirical evidence of the relationship between education and inter-regional migration focuses primarily on location choices of recent college- or university graduates <references>. We will elaborate more on the findings of this literature below. However, to our knowledge, only very few studies have described the location decisions among young adults, when they move away from their parents. No previous study has, to our knowledge, described the causal effect of high school on migration among young adults even though this is by far the most geographically mobile group. Therefore, policy makers still lack evidence about how local education policies affects migration and the regional population distribution. The aim of our paper is to contribute toward filling this gap.

Previous research of inter-regional migration has primarily focused on the adult population in the labour force, describing which local factors that attract and maintain highly skilled workers in a region ((Détang-Dessendre, Goffette-Nagot, & Pigué, 2008; Mellander, Florida, & Stolarick, 2011). Several empirical studies describe inter-regional migration patterns in the transition between education and work. More specifically the studies examine the location choices of recent university graduates and find that most of the university graduates stay in the city where they completed their education. Using US panel data, (Kodrzycki, 2001) find that 70 percent of the students live in the city where they graduated, five years after graduation. In addition (Busch & Weigert, 2010; Haapanen & Tervo, 2012) apply duration models to describe the probability of migration during education and up

⁸ In the following econometric analysis, we follow youth from they are 18 and until they are 24 to see if they complete high school and migrate out of a rural or non-urban area.

until 10 years after graduation. Busch and Weigert (2010) confirm the results from (Kodrzycki, 2001) using German data and show that 70 percent of the university graduates live in the university region 10 years after graduation. In addition, they show that a third of the outmigration occurred within the first year after graduation, and after five years the probability of moving was almost non-existent. In sum, the empirical evidence suggests very limited geographic mobility among youth, after graduation from a higher tertiary education.

Very few empirical studies examine youth migration, which may be surprising considering that this is by far the most mobile age group. One explanation of this gap in the literature, could be the lack of available data. Berck et al. who provide useful evidence on Sweden, is an exception. Using Swedish register data, (Berck et al., 2016) estimate the education and location choice of young adults in Sweden, as a function of characteristics of the region they might migrate to. The results are consistent with investment into further education being associated with preferences for high per capita tax bases.

4. The Danish education system

This section describes the secondary Danish educational system and the requirements for admission to a higher education. The purpose is to clarify the educational choices that young adults face after completing compulsory education in Denmark.

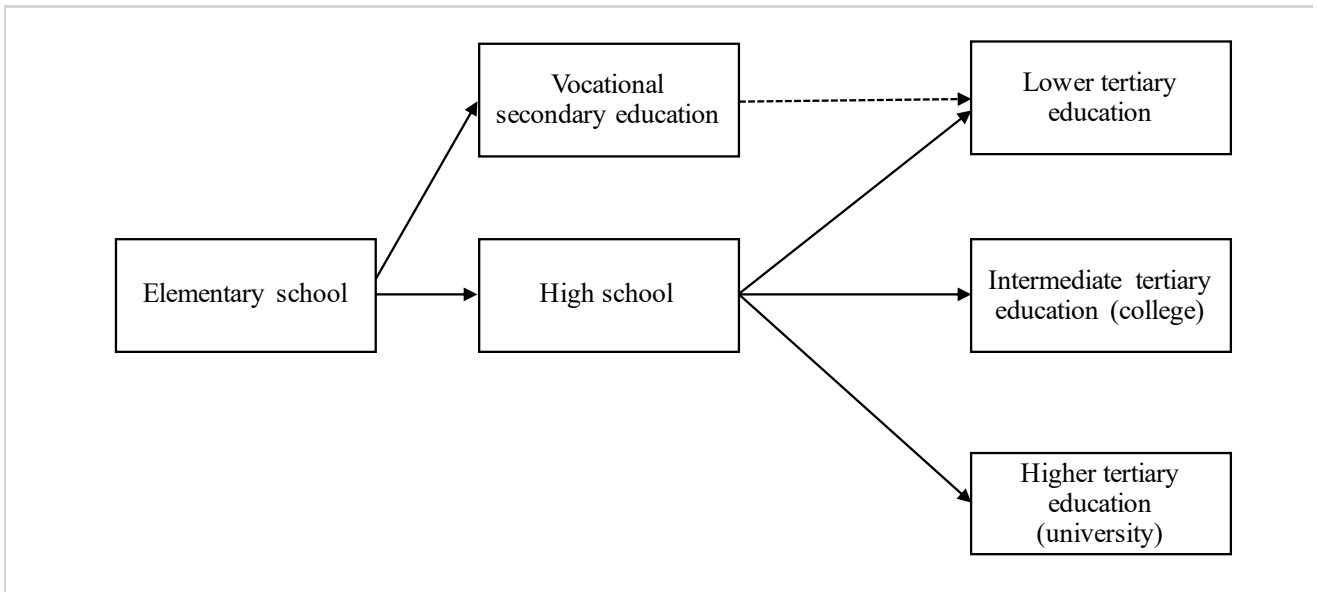
The education system in Denmark is universal. There are no tuition fees in secondary or in higher educations. Compulsory education in Denmark consist of primary and lower secondary elementary education, from grade 0 (age 5–6) to grade 9 (age 15–16). After 9th grade, further education is voluntary. Pupils can choose to leave the educational system, continue in 10th grade or enter upper secondary school. Students that do not directly enrol in upper secondary school may enter later, with no loss of rights or opportunities for enrolment.

The upper secondary school comprises two main tracks of education: high school and vocational secondary education. Most student (about 90 percent of a birth cohort) eventually choose one of the two tracks. High school consists of academic tracks, such as mathematics, technical studies and linguistics. Vocational upper secondary education consists of occupation-specific tracks such as carpentry, bricklaying, mechanics and hairdressing. The vocational system is a dual educational system such that students shift between school-based learning and practical apprentice training. The two types of education are placed in different institutions and are usually not located on the same geographical location.

Tertiary education in Denmark is on three levels: lower tertiary education (LTE), intermediate tertiary education (ITE) and higher tertiary education (HTE). A formal requirement for admission to all

tertiary education is usually a high school degree⁹. The system is depicted in Figure 3 below, ignoring the fraction of students that entirely drop out of the educational system. Figure 3 illustrates that, vocational education is a “dead end” in terms of the option of further education, while high school offers the option of entering further education.

Figure 3: Typical pathways in the Danish educational system



5. Data

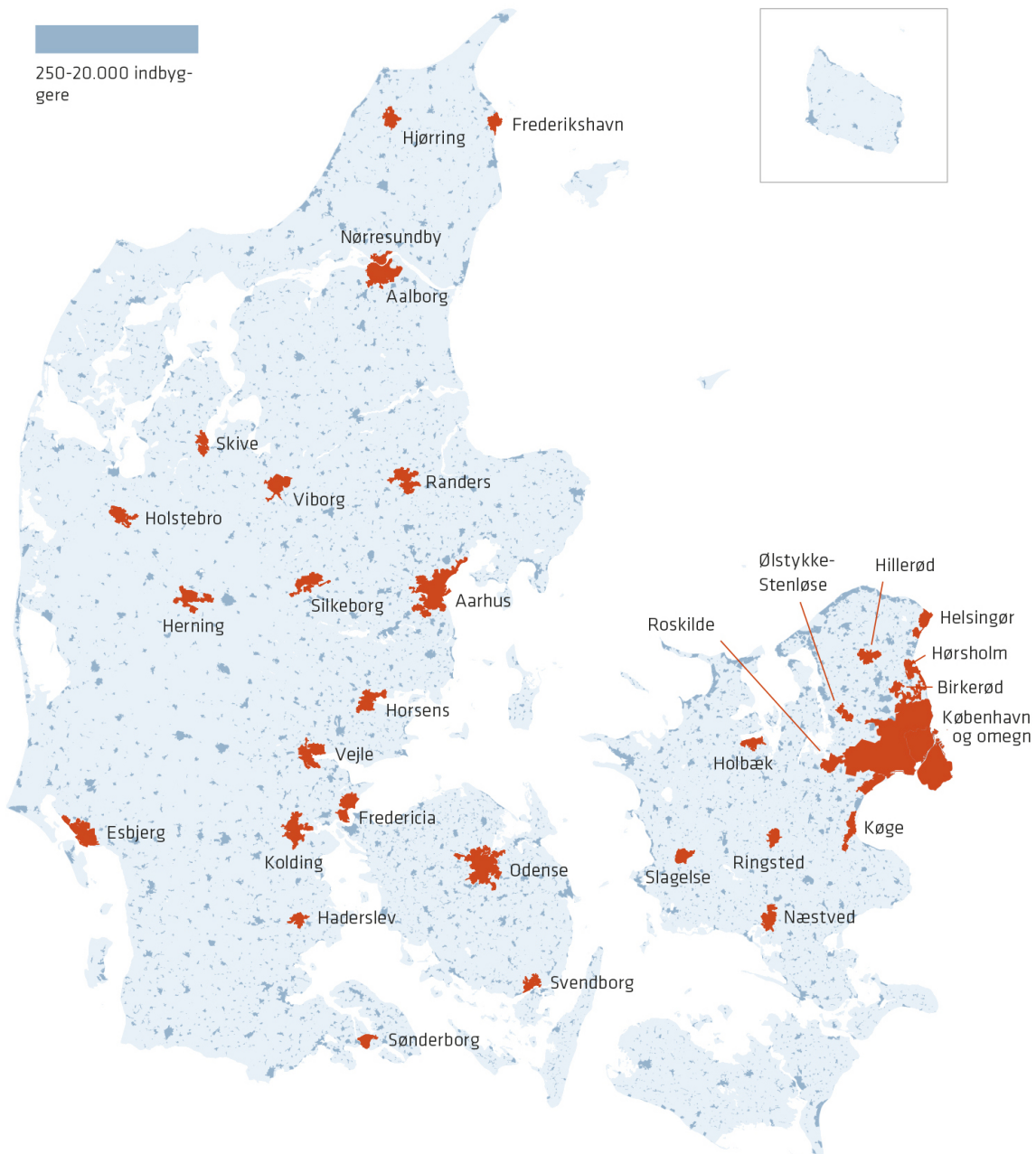
We use panel data from the population registers of Statistics Denmark in combination with detailed distance measures between place of residence and place of nearest high school. The registers contain rich information about individual characteristics, such as educational choice, change of residence, family members, place of residence and income. The sample encompasses five different birth cohorts: all individuals born from 1986 – 1990, who at age 15 were resident in a rural area in Denmark. Furthermore, we restrict our sample to young individuals who stayed in a rural area until they turned 18. We do this to ensure that the individuals we observe do not move before graduating from high school. Figure A1 in appendix illustrate our sampling strategy.

A rural area is defined as rural areas and towns with less than 20.000 inhabitants in 2016. We hold the city-area fixed back in time to avoid that changes in city borders can affect the results. This

⁹ However, some lower- and intermediate tertiary tracks may sometimes allow students to use selected vocational secondary education to meet the admission requirement

means that the 33 largest cities in Denmark is defined as “urban areas” in our analyses. Figure 4 below shows the names of the urban areas and where they are located, marked with red. Our sample represents a significant share of Danish youth as 60 percent of the 15-years old resided in a rural area, during the time period we analyse.

Figure 4: Map over rural and urban areas in Denmark in 2016



Definition of inter-regional migration

We define inter-regional migration as a change of residential address from a rural area to an urban area. The dependent variable is a dummy variable, which is equal to 1 if the individual has moved to an urban area, before turning 25. Only moves away from the childhood home, are counted as moves. This means that individuals who moves to the same residential address as their parents', are not defined as movers. We look at repeated moves from the age of 18 years until the age of 25 years. If just one of the moves goes towards an urban area, the dependent variable is equal to 1.

We make robustness checks to our definition of urban area, see table A3 in appendix. We run three robustness checks: One, where we redefine an urban area as one of the four largest cities in Denmark (Copenhagen, Aarhus, Aalborg and Odense), second, we define an urban area as cities with at least one university. Lastly, we rerun the analysis where we only look at moves to the Capital area, Copenhagen. The first two definitions of urban areas, do not change the results substantially, see table A3 in appendix.

Explanatory variables

The explanatory variables measure individual characteristics, family characteristics and attributes of location and distance to nearest large city at age 15 years. The only exception is our primary explanatory variable measuring if the individual complete high school or not, before turning 25. "High school" is a dummy variable, which is equal to 1, if the individual has graduated from high school before turning 25 years. If the individual has completed a vocational education or has no secondary education the variable is equal to 0. About 12 % of the sample completes both a vocational and a high school education, before turning 25. In these cases, we keep the first completed upper secondary education.

Table 1: Definitions of variables

Variables	Variable names	Description
Dependent variable	Move	Dummy variable equal to 1 if moved to a large city, before age 25
Individual characteristics		
Education choice	High school	Dummy variable, equal to 1 if completed high school before age 25
Gender	Female	Dummy variable, equal to 1 if female
Immigration status	Danish	Reference category
	First generation immigrant	Dummy variable, equal to 1 if first generation immigrant
	Second generation immigrant	Dummy variable, equal to 1 if second generation immigrant
Year of birth	1986	Reference category
	1987	Dummy variable, equal to 1 if born in 1987
	1988	Dummy variable, equal to 1 if born in 1988
	1989	Dummy variable, equal to 1 if born in 1989
	1990	Dummy variable, equal to 1 if born in 1990
Grades in primary school	Written Danish	Exam grade in written Danish in 9th grade
	Missing grade in Danish	Dummy variable, equal to 1 if missing exam grade
	Written Math	Exam grade in written Math in 9th grade
	Missing grade in Math	Dummy variable, equal to 1 if missing exam grade
Geographical characteristics		
	Municipality	Residential municipality at age 15
	Geographical region	Residential region at age 15
	Distance to large city	Distance between residential address and nearest large city (km)
	Distance to school	Distance between residential address and nearest high school (km)
Family characteristics		
Mother's education	Missing education information	Dummy variable, equal to 1 if education information is missing
	Elementary school	Reference category
	High School	Dummy variable, equal to 1 if high school
	Vocational	Dummy variable, equal to 1 if vocational education
	Short-cycle higher education	Dummy variable, equal to 1 if short-cycle higher education
	Medium cycle higher education	Dummy variable, equal to 1 if medium-cycle higher education
	Long-cycle higher education	Dummy variable, equal to 1 if long-cycle higher education
Father's education	Missing education information	Dummy variable, equal to 1 if education information is missing
	Elementary school	Reference category
	High School	Dummy variable, equal to 1 if high school
	Vocational	Dummy variable, equal to 1 if vocational education
	Short-cycle higher education	Dummy variable, equal to 1 if short-cycle higher education
	Medium cycle higher education	Dummy variable, equal to 1 if medium-cycle higher education
	Long-cycle higher education	Dummy variable, equal to 1 if long-cycle higher education
Household income	Household income (log)	Log of yearly household income (in DKR).
	Poor	Dummy variable, equal to 1 if in the poorest percentile
	Rich	Dummy variable, equal to 1 if in the richest percentile

Note: All attributes were measured when individuals were 15 years old, if nothing stated

Distance to school

Our instrumental variable is distance to nearest high school (we return to further description of the assumptions behind the IV approach in the empirical strategy section). Distance to school, is measured as the distance from the young person's residential address to the nearest high school¹⁰. We measure the distance through road network in year the person turned 15 years old, using GIS-software. Young adults living on small islands, without a bridge to the mainland, are excluded from the analysis. In total 5 percent of the sample is excluded on this behalf.

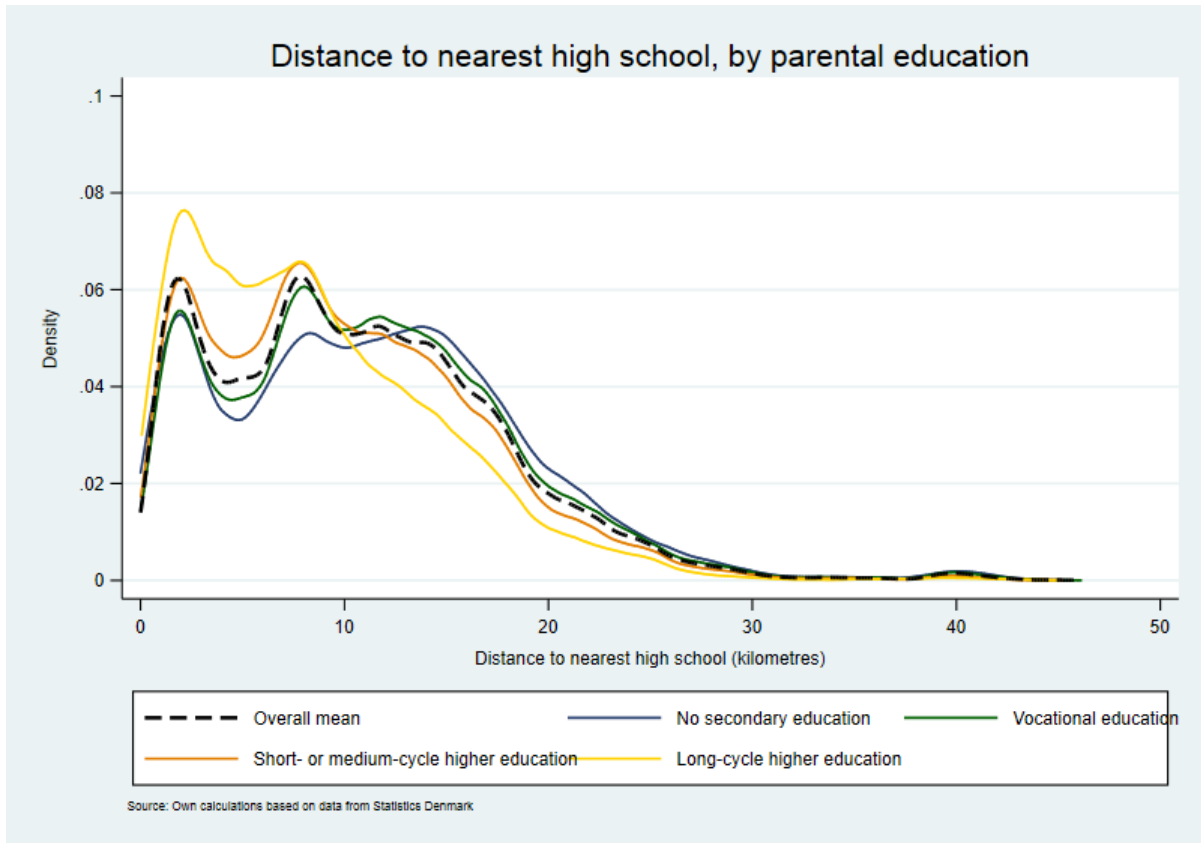
The average distance to high school is 10 kilometres, with a minimum distance of 0 km and a maximum distance of 46 kilometres to the nearest high school¹¹. This may seem like relatively long distances for a small country like Denmark. However, keeping in mind that our population of young adults live in smaller cities or rural areas, this seems reasonable. Figure 5 shows the distance distribution, split by parental education. The black dashed line illustrates the overall mean. The figure illustrates that most of the population have access to a high school within a relatively short distance; 50 percent have less than 10 kilometres to the nearest school and about 25 percent have less than 5 kilometres (Table A1 in appendix shows the cumulative distance distribution in 5-kilometre intervals).

When we stratify the distance distribution by socioeconomic background (SES), we see that young adults, where one or both parents hold a long-cycle education (the yellow line) on average live closer to high schools. At the same time, low SES students, where none the parents have a secondary education (green and blue line), have, on average, longer distances to school. This indicates that distance i.e. our instrumental variable, is correlated with observed parental education, as you would expect. We return to this potential endogeneity problem in our presentation of results and robustness checks.

¹⁰ To calculate the distances for the individuals, we use the exact geographic co-ordinates for each of the identified educational institutions and the co-ordinates for the bottom-left corner of the geographic defined quadrant 100 x 100 meter in size in which the residence of the student was located.

¹¹ The families have to arrange and pay for their own transportation to high schools, as there are no school buses provided. However the public transportation system is relatively well-developed in Denmark and the prices are subsidized by the state.

Figure 5: Distance to nearest high school by parental education



In table 2 we present descriptive statistics of the variables in the analysis. About 71 percent of the sample moved to an urban area before turning 25 years, and about 55 percent completed high school. The relatively low completion rate (from an international perspective) reflects the structure of secondary educations in Denmark, where a large share completes vocational secondary education. About 47 percent of the sample are females and 96 percent are non-immigrants. We have information about each individual's exam grade in 9th grade in written Danish and math. The grade point average in 5.5 in written Danish and 5.3 in Math, which is a bit lower than the national average at 6.35 and 6.2 respectively¹². The majority of the sample lived in the regions of Central Jutland (27 percent) or Southern Denmark (28 percent). Only about 9 percent grew up in the Capital Region, because there are relatively few rural areas in the Capital Region. The mean distance to the nearest high school is 10.7 kilometres where the mean distance to the nearest large city is about 26 kilometres. We also take parental education and household income into account, in our analysis. The distribution of mother's and father's highest education is also shown in Table 2. About one-fourth of the mothers

¹² Information about the national GPA average comes from the Ministry of Education from 2012

<https://uddannelsesstatistik.dk/Pages/Reports/1809.aspx>

and fathers respectively, have no further education past elementary school. 39 percent of the mothers and 46 percent of the fathers hold a vocational education. This means that only about 30 percent of the mothers and 20 percent of have a further education at the tertiary level. The education distribution reflects both the age cohort of the parents (typically born in the 1950's and 1960's) and the parents live in rural areas. The mean annual disposable household income

Table 2: Individual- and family specific variables

Table 2: Individual- and family specific variables

Variables		Mean	Std. Dev.	Min	Max
Dependent variable	Moved to an urban area	0.71	0.45	0	1
Individual characteristics					
Completed high school		0.54	0.50	0	1
Gender	Female	0.47	0.50	0	1
	Male	0.53	0.50	0	1
Immigration status	Danish	0.96	0.20	0	1
	First generation immigrant	0.03	0.17	0	1
	Second generation immigrant	0.01	0.10	0	1
Year of birth	1986	0.19	0.39	0	1
	1987	0.19	0.39	0	1
	1988	0.20	0.40	0	1
	1989	0.20	0.40	0	1
	1990	0.22	0.41	0	1
Grades in primary school	Written Danish	5.46	3.04	-3	12
	Dummy for missing grade in Danish (1=missing)	0.08	0.27	0	1
	Written Math	5.29	3.19	-3	12
	Dummy for missing grade in Math (1=missing)	0.08	0.27	0	1
Geographical characteristics					
Geographical region	Northern Jutland	0.16	0.37	0	1
	Central Jutland	0.27	0.44	0	1
	Suthern Denmark	0.28	0.45	0	1
	Capital Region	0.09	0.29	0	1
	Zealand Region	0.20	0.40	0	1
Municipality	Municipality Fixed Effects				
Distance	Distance to high school (in km)	10.67	6.87	0	46
	Distance to large city (in km)	26.05	15.37	1	88
Family characteristics					
Mother's education	Missing education information	0.02	0.15	0	1
	Elementary school	0.26	0.44	0	1
	High School	0.04	0.20	0	1
	Vocational education	0.39	0.49	0	1
	Short-cycle higher education	0.04	0.19	0	1
	Medium cycle higher education	0.22	0.41	0	1
	Long-cycle higher education	0.03	0.18	0	1
Father's education	Missing education information	0.06	0.24	0	1
	Elementary school	0.23	0.42	0	1
	High School	0.03	0.17	0	1
	Vocational education	0.46	0.50	0	1
	Short-cycle higher education	0.04	0.20	0	1
	Medium cycle higher education	0.11	0.31	0	1
	Long-cycle higher education	0.06	0.23	0	1
Household income (Log)	Household income	12.79	0.38	5.6	17.1
	Poorest 1%	0.01	0.10	0	1
	Richest 1%	0.01	0.10	0	1

N = 146,375

6. A theoretical model

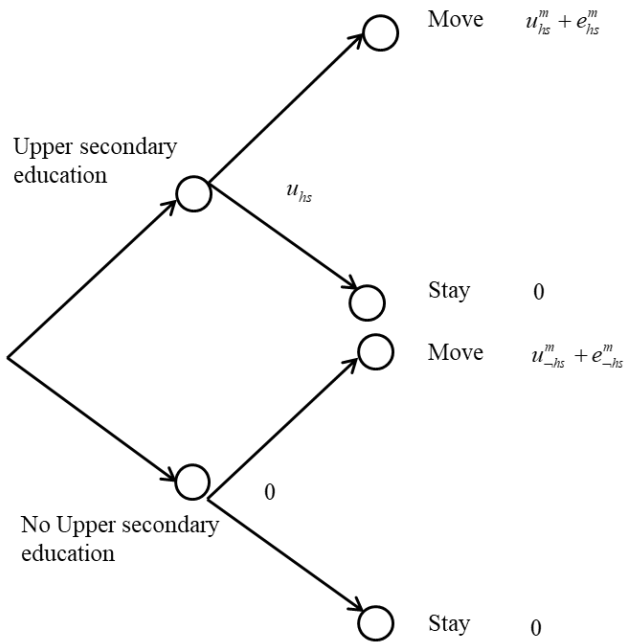
In this section, we present a theoretical model of student choice of completing high school (HS) and subsequent choice of migrating from rural to large urban areas (Move). The section illustrates first a potential mechanism behind the causal effect of HS on the decision to leave a non-urban area.

Second, our theoretical model illustrates the endogenous aspect of choosing HS.

In the model the student and his or her family is faced with two consecutive choices. First, whether to complete high school and second, whether to move out of the rural area. For each choice there is instantaneous utility associated with the choice. For the completion of high school this is u_{hs} . Utility for not taking high school (including enrolling and dropping out) is normalized to zero. The instantaneous utility of completing HS may consist of the utility of adhering to family norms and the cost of studying. u_{hs} may therefore be either positive or negative.

If the student decides to move after completing high school the instantaneous utility contains two components, one which is known when completing high school, u_{hs}^m and a component that is first realized after moving, e_{hs}^m . If the student decided not to complete high school and decides to move, utility likewise contains two components, one which is known when deciding not to enroll into high school, u_{-hs}^m and a component that is first realized if moving, e_{-hs}^m . The instantaneous utility if the students decided not to move is normalized to zero. Special attention should be devoted to the two unknown terms realized after moving, e_{hs}^m and e_{-hs}^m . From the time before entering HS these are option values. We think of the option values of HS, over and above the utility of working in an urban labor market encompasses the option value of entering tertiary education. This would involve long commuting time if staying in a rural area. The situation is illustrated in figure 6 below.

Figure 6: Choice behavior



When the student must decide whether to enter high school he has to take into account not only the instantaneous utility from either entering or not entering high school, but also the expected maximum utility from either moving or not moving given the choice of high school. Upon assuming that the e 's are standard normal distributed, the expected maximum utility from entering and completing high school is

$$V_{HS} = \underbrace{u_e}_{\substack{\text{Instantaneous} \\ \text{Value of} \\ \text{completing} \\ \text{US}}} + E \max \left[\underbrace{u_{hs}^m + e_{hs}^m, 0}_{\substack{\text{Expected value of being} \\ \text{able to make the optimal} \\ \text{choice between} \\ \text{moving versus} \\ \text{staying with US}}} \right] = u_e + \underbrace{u_{hs}^m \Phi(u_{hs}^m)}_{\substack{\text{Expected value} \\ \text{of moving}}} + \underbrace{\varphi(u_{hs}^m)}_{\substack{\text{Option value} \\ \text{of being able} \\ \text{to make the} \\ \text{optimal choice} \\ \text{between} \\ \text{moving and} \\ \text{staying}}}$$

where $\Phi(\cdot)$ and $\varphi(\cdot)$ are the standard normal cdf and pdf respectively. Similarly, we find the expected value of not completing high school:

$$V_{-HS} = E \max \left[u_{-hs}^m + e_{-hs}^m, 0 \right] = u_{-hs}^m \Phi(u_{-hs}^m) + \varphi(u_{-hs}^m).$$

Now the choice of completing is thus:

$$V_{HS} > V_{-HS}$$

\Leftrightarrow

$$u_e > \underbrace{u_{-hs}^m \Phi(u_{-hs}^m) - u_{us}^m \Phi(u_{hs}^m)}_{\text{Differences in expected value of moving}} + \underbrace{\varphi(u_{-hs}^m) - \varphi(u_{hs}^m)}_{\text{Difference in option value of moving}}.$$

From this we see that to choose high school either the expected value and/or the option value of moving with HS must outweigh the similar values without HS and the difference must justify the cost of completing high school. Thus, if u_{hs} is negative the net difference of moving with HS relative to non-HS must be positive.

Once HS is completed, moving with HS is: $e_{hs}^m > u_{hs}^m$ and similarly for moving without HS. From the perspective of the student, the choice of HS is a deterministic choice whereas the choice to move is probabilistic at the time of making the choice of HS. For the econometrician, however, both choices may have to be modeled as probabilistic if components of the choices are unobserved (both components known to the student and components not known). Further, in this case, the choice of HS may, from the perspective of the econometrician, depend on unobserved components stemming the decision to move and hence the two components must be estimated simultaneously. We note that moving with HS entails the opportunity to enter tertiary education, hence a relatively “large” option value. Moving without HS entail only the opportunity to enter skilled or unskilled occupations, a relative “low” option value because the labor market conditions for unskilled and skilled workers may not differ much from rural to urban areas. Hence option value of HS will in many cases be larger than option values without HS. Option value may therefore be an important driver of the choice of HS and a major value of HS is the option of entering tertiary education in the future.

7. Empirical strategy

In this section, we describe our empirical strategy. As described in the previous section, it is very natural to assume that the decision to move and the decision to complete US shares the same unobserved characteristics and therefore, that from the perspective of the econometrician, the decision to complete US is an endogenous variable.

Our empirical model is the linear probability model:

$$P(y_i = 1 | H_i, \mathbf{x}_i) = \alpha + \delta H_i + \boldsymbol{\beta} \mathbf{x}_i + \varepsilon_i \quad (1),$$

where $\alpha, \delta, \boldsymbol{\beta}$ are parameters to be estimated. y is a binary endogenous variable, equal to one if the individual moves out of a rural area and into an urban area before the age of 25 and 0 if the individual stays in a rural area. \mathbf{x} is a vector of exogenous control variables and H is a potential endogenous variable indicating if the individual has completed high school. It takes the value one if high school is completed and zero otherwise. ε is an error term capturing the effect of omitted variables. The endogeneity of H comes from the plausibility of the correlation between completing high school and unobserved characteristics of the individual.

To circumvent the endogeneity of H we employ an instrumental variable approach.

That is, we introduce an auxiliary regression equation stating that completing high school depends on an instrumental variable, distance to nearest high school, that only affects completing high school but not whether one moves to an urban area. The first stage equation is

$$P(H_i = 1 | x_i, z_i) = \boldsymbol{\theta} \mathbf{x}_i + \gamma z_i + \dot{U}_i \quad (2),$$

where z is the instrumental variable and where we expect \dot{U} and ε to be correlated due to common omitted variables. However, one may argue that parents locate near or far from high schools in response to expectations and aspirations of their children completing high school, such that parents with high expectations locate closer to high schools than parents with lower expectations. If parental expectations and hence distance is correlated with \dot{U} our exclusion restriction does not generally hold. We return to this issue below.

In the case the treatments effect, δ , is heterogeneous, the interpretation of the 2SLS or Wald estimate of δ is then the local average treatment effect (LATE), see (Imbens & Angrist, 1994). This means,

that we estimate the treatment effect for the compliers (those who are affected by distance and choose high school if it is located close to their home, and not if the distance is longer.

Robustness check on the exogeneity assumption of the instrumental variable

Our instrumental variable is distance to HS. In studying return to college education many researchers have used a similar instrumental variable, distance to college, e.g. (Cameron & Taber, 2004; Card, 1995; Carneiro, Heckman, & Vytlacil, 2011; Currie & Moretti, 2003).

However, our identification strategy relies on the assumption that, conditional on x , distance is independent of U and ε . If some parental and student characteristics are unobserved but relevant in both equations in the system (1) and (2), then distance being independent of U and ε amounts to assuming that unobserved parental and student characteristics are independent of distance to high school. This may not be a tenable assumption, e.g. (Cameron & Taber, 2004) who shows that distance to college is correlated with cognitive ability.

We only have an imperfect measure of this variable in terms of selected school grades. We have run our first stage regressions with and without control variables (x) and it is evident that the estimated effect of distance on high school is sensitive to whether controls are included in the first stage. This indicates that distance is correlated with observed controls and thus makes a case for distance also likely to be correlated with the both U and ε . This violates a central assumption behind the use of distance as an instrumental variable. To rectify the problems with our instrumental variables we employ robustness checks. We use changes in distances to high school institutions. Some new institutions opened up while others close down during our observation period. This implies that distance may change between siblings in the same family with the same location. Under the assumption that parents cannot predict whether a new high school facility opens up in their area change in distance between siblings will arguably be exogenous. We exploit the fact that we can link individuals to their parents and their siblings in the data. We then study siblings, who lived the same place, with the same parents, but are exposed to different distance to education, because a new institution opened up closer to them between the siblings 15 years birthday. Change in distance across siblings will most of the time in our application use a more restricted number of openings and closures of facilities as the usual age span between siblings is less than five years in our data. Our approach amounts to a sibling fixed effect IV approach.

We also employ a sibling and twin fixed effect estimation. We decompose the error term in (1) ε into two terms, a family fixed effect that reflect family aspirations, genetic endowments that benefits high school completion and that may be correlated with completing high school and an individual component that is assumed independent of completing high school. The latter assumption may be questionable, and we test it by adding important individual level characteristics. such as comprehensive school grades and by restricting our estimation sibling fixed effects to same sex twins (who on average share more individual characteristics, assuming the remaining unobserved individual characteristics independent of high more tenable). All though twin studies have a mixed reputation in economics, see (Goldberger, 1979; Heckman, 2007),¹³ we believe that twin FE's is a viable and interesting way of gauging average effect.

8. Results

In this section, we first present our results from the instrumental variables regression using distance to high school as instrument and then the fixed effect estimates investigating differences between siblings and same sex twins.

We begin our reporting of the instrumental variables regressions with the first stage regression of distance to high school institutions and next we present the second stage results of the effect of high school on the decision to move out of a rural and into an urban area. We present our second stage results alongside with linear probability model (LPM) results to illustrate the difference between our raw LPM estimates and the LATE estimates. We further present results stratified on socio economic background (SES) of the student because we expect that our instrument has a larger fraction of compliers (students whose decision to enroll in high school is affected by the IV) among low SES students as opposed to higher SES students.

In table 3 we present results for our first stage results. In the first column, we present first stage results without any additional covariates, in the second column, we present a model with a limited set of covariates (student characteristics) and in the third column, we present a model with a full set of covariates (parental characteristics). We add covariates gradually to see if adding covariates changes the effect of the instrument. This would give an indication of whether the instrument is correlated with observables and hence if it is likely that it is also correlated with unobservables.

¹³ However, see (Royer, 2009) or (Cesarini, Dawes, Johannesson, Lichtenstein, & Wallace, 2009) for recent applications of twin fixed effect in economics.

Table 3: Results of 1. stage regression

Y = Completed high school at age 25			
	First stage - covariates	First stage - reduced model	First stage - full model
Variable	coefficient / SE	coefficient / SE	coefficient / SE
Distance to high school institution, at age 15	-0.0103*** (0.0005)	-0.0074*** (0.0004)	-0.0051*** (0.0004)
Distance to high school institution, squared at age 15	0.0002*** (0.0000)	0.0002*** (0.0000)	0.0001*** (0.0000)
Intercept	0.6264*** (0.0032)	0.0606*** (0.0046)	-1.1550*** (0.0490)
Individual covariates		YES	YES
Parental covariates			YES
Municipality FE			YES
Number of observations	146,375	146,375	146,375
R2	0.0066	0.3855	0.4136
F	483.71	303.62	110.73

note: *** p<0.001, ** p<0.01, * p<0.05

From the table we first see that our instrument has a strong significant effect on our endogenous variable, the decision to complete the academic track of high school. Second, we see that the effect of the instrument change over models with additional covariates. Both adding individual and parental covariates as well as municipality fixed effects changes the effect of the instrumental variable in the first stage regression. This indicates that the instrument is correlated with observed parental characteristics, which again makes it somewhat harder to maintain the important assumption that it is uncorrelated with unobserved parental characteristics, i.e. that the exclusion restriction is viable. We return to this in our robustness check. For the moment, we proceed to second stage results maintaining the assumption that the IV is uncorrelated with relevant unobserved variables.

In table 4 we show our second stage results together with LPM results. From the table we see that our second stage results, which are local average treatment effects (LATE), are strongly significant and indicates a very large effect of high school on the decision to move out of a rural area and into an urban area. The probability of moving increases with 66 percent if one has a high school

degree compared to if one has not. The estimate up from 26 percent for the LPM model with a full set of covariates.

Table 4: Estimation results

Dependent variable: Moving to a large city, before turning 25		
	LPM	LATE
Variable	coefficient / SE	coefficient / SE
Completed high school at age 25	0.2641*** (0.0027)	0.6587*** (0.0750)
Intercept	-0.1603*** (0.0512)	0.3132*** (0.1052)
Number of observations	146,375	146,375
R2	0.2341	0.1245
F-test (IV)		110.7
Individual controls	YES	YES
Family control	YES	YES
Municipality FE	YES	YES
IV	NO	YES

note: *** p<0.001, ** p<0.01, * p<0.05

The reason that there is such a large difference between the LPM and the LATE may be rooted in the difference between the effect for the compliers (LATE) and average treatment effect (of which the LPM is a biased estimate of). Those affected by distance in their choice of high school have a low probability of moving without a high school degree and their decision to move is very much affected by completing a high school degree while non-compliers move to urban areas with or without a high school degree. This finding is supported by our findings for heterogenous effects below.

Heterogeneous effects

To see if there is effect heterogeneity across the population in our study, we split our sample according to parental education. Parental background is an important driver for educational choice. Hence, we expect compliers to vary across parental background. Studying heterogenous effects across parental background may shed light on the mechanisms behind our large overall complier effect in table 4. In table 5 below we show how the sample is split by parental education.

Table 5: Population split in four groups by parents' highest education

Groups:	Parents' highest education	Frequency	Percent
SES 1:	Both parents have no education beyond primary school	21,084	14%
SES 2:	One or both parents hold a vocational education	71,141	49%
SES 3:	One or both parents hold a shorter- or medium cycle higher education	43,302	30%
SES 4:	One or both parents hold a long-cycle higher education	10,848	7%
Total		146,375	100%

Table 6 below show the distance distribution split by parental education.

Table 6: Distance to high school by parental education

Groups:	Distance to nearest high school (IV)	Mean	Std. Dev.	Min	Max
SES 1:	Both parents have no education beyond primary school	11.43	7.29	0	45.15
SES 2:	One or both parents hold a vocational education	11.11	6.96	0	46.12
SES 3:	One or both parents hold a shorter- or medium cycle higher education	10.1	6.55	0.01	45.23
SES 4:	One or both parents hold a long-cycle higher education	8.67	6.18	0.06	45.04

From the table we see that SES group 1 of parents with no secondary education makes up 14 percent of the sample, while parents with a vocational education, SES group 2, make up almost half the

sample. Parents with long cycle higher education, SES group 4, is only seven percent of the sample¹⁴. The skew educational distribution towards relatively short-term educations of parents reflects both the age group of the parents (typically born in the 1950's and 1960's) and the fact that the parents constitutes a rural population.

In table 7 we show first stage results by SES groups (we show similar results for the LPM in the appendix.).

Table 7: First stage regressions split by parental education

Y = Completed high school at age 25				
	First stage: SES1	First stage: SES2	First stage: SES3	First stage: SES4
Variable	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE
Distance to nearest high school at age 15	-0.0063*** (0.0010)	-0.0057*** (0.0006)	-0.0043*** (0.0008)	-0.0017 (0.0014)
Distance to nearest high school, squared, at age 15	0.0002*** (0.0000)	0.0001*** (0.0000)	0.0001*** (0.0000)	0.0000 (0.0001)
Intercept	-0.7071*** (0.1132)	-1.3444*** (0.0760)	-1.1699*** (0.0948)	-0.3994*** (0.1309)
Number of observations	21,084	71,141	43,302	10,848
R2	0.3650	0.3704	0.3296	0.2430
Individual controls	YES	YES	YES	YES
Family control	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES

note: *** p<0.001, ** p<0.01, * p<0.05

We find strong and significant first stage results for the first (least educated) SES group. Further, the effect of distance is declining across SES groups, such that the first stage is strongest for SES group one and completely absent in SES group four. Hence, the weaker the educational background of the parents, the stronger the first stage effect.

In table 8 we show second stage results split by SES groups.

¹⁴ We divide the sample according to the highest education of the two parents. This means that at least one of the parents hold a vocational education in SES group 2, that at least one of the parents holds a university degree in SES group 4 and so forth.

Table 8: Results of 2. stage regression split on parental education

Dependent variable: Moving to a large city before turning 25				
	IV: SES1	IV: SES2	IV: SES3	IV: SES4
Variable	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE
Completed high school at age 25	0.7936*** ▲ (0.2066)	0.6647*** ▲ (0.1023)	0.4285*** ▲ (0.1260)	0.9376 ▲ (0.7253)
Intercept	0.4551** ▲ (0.2143)	0.3696** ▲ (0.1672)	0.0319 ▲ (0.1765)	0.5558* ▲ (0.3336)
Number of observations	21,084 ▲	71,141 ▲	43,302 ▲	10,848 ▲
R2	0.0205 ▲	0.0879 ▲	0.1625 ▲	-0.4891 ▲
Individual controls	YES	YES	YES	YES
Family control	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES
IV	YES	YES	YES	YES

note: *** p<0.001, ** p<0.01, * p<0.05

From the table, we find declining effects of completing high school across parental education. For SES group 1 (unskilled parents) we find an effect size of almost 80 percentage points. The effect size drops to 66 percentage points for SES group 2 and 43 percentage points for SES group 3 and its insignificant for SES group 4 because there is no first stage for this group. This is properly because the fraction of compliers in SES group 4 is very small. We conjecture that students with very well-educated parents are, to a large extent “always takers” and hence take high school irrespective of the distance to high schools, while the lower SES groups are more inclined to take high school if cost, including transportation cost is low.

In sum, we find large significant complier effects of completing high school on the probability to move from a rural area to a larger city. However, before we proceed to conclusions, we conduct some robustness checks. More specifically we address the potential endogeneity of our IV by refining the IV to look at differences in distances to high schools arising from opening of new high school institutions after the parents of the students moved to the residential area where they lived when the student was 15.

Robustness checks

In this section we employ several robustness checks. First, we use a refined version of our instrumental variable, distance to high school. Utilizing that between siblings (when they are age 15) distance to nearest high school may change due to openings and closing of institutions we get within sibling variation in distance to nearest high school which may arguably be independent from parental background characteristics and also student characteristics. This leads to a fixed effect instrumental variable (FE-IV) approach (Wooldridge, 2005). Because there is only a limited number of openings and closings of high schools, we run the first stage as a binary indicator of whether the distance is more (=1) or less (=0) than 7 kilometers. Hence, the FE-IV first stage measures whether the distance changes to more/less than 7 kilometers between siblings. In table 9 below, we show first stage results using the change in distance to high school between siblings.




Table 9: Results of 1. stage regression for sibling IV Fixed effects

Endogenous variable: Completed high school at age 25	
	First stage - full model
Variable	coefficient / SE
Distance to nearest high school. Dummy = 1 if distance exceed 7 kilometres	-0.1572*** █ (0.0601)
Intercept	█ 0.2817 █ (0.1925)
Individual covariates	YES
Parental covariates	YES
Municipality FE	YES
Number of observations	█ 41,174
R2	█ 0.2333
F	6.85

note: *** p<0.001, ** p<0.01, * p<0.05

From the table we find that when the distance to the nearest high school change more than 7 kilometers between siblings, the probability of attending high school drops 16 percent. The effect is significant at the 0.1 percent level. However, the F statistic is not above the recommended value of 10, see (Bound, Jaeger, & Baker, 1995). Hence, we proceed with caution when inspecting our LATE based on sibling fixed effect IV estimation. We return to this below in the comments to table 9 which shows the second stage results for the sibling fixed effect IV.

Table 10: Sibling IV fixed effects model

Dependent variable: Moving to a large city before turning 25	
Sibling IV FE	
Variable	coefficient / SE
Completed high school at age 25 	0.8530* (0.4554)
Intercept 	-0.1863** (0.0818)
Number of observations 	41,174
IV	YES

note: *** p<0.001, ** p<0.01, * p<0.05

From table 10 we find a significant effect of high school completion on the probability of moving out of rural area of 85 percent, which is significant at the 5 percent level. Our instrument is weak (F-statistic less than 10). However, our sibling fixed effect IV gives a comparable (85 percent versus 79 percent) results as our stratified main IV regression for the lowest SES group. We take two things from this. First, the very large effects from the stratified (by SES) IV analysis seems in line with our more robust but much less efficient siblings fixed effect IV analysis and hence add credibility to our main specification (IV using distance). Second, the compliers in our sibling fixed effects analysis seems to be similar to the compliers in the lowest SES group. In sum, we find support to a causal interpretation of our main IV analysis from our sibling fixed effect IV analysis.

As a final analysis to investigate the overall effect of high school on the probability to move out of rural areas we show the results of sibling and same sex twin fixed effects analysis. To identify twins, we constructed a smaller dataset containing solely siblings, who were born on the same date and have the same sex. We keep twins that share residential address, which means that they have equal distance to school and have had the same upbringing. The same sex twin fixed effect (FE) estimate purges all family fixed effects and to the extent that same sex twins share genetic make up, also (some of the) genetic effects. We show our sibling and same sex twin fixed effect results in table 10 below.

Table 11: Result of robustness analysis on siblings and same sex twins

Dependent variable: Moving to a large city before turning 25		
	Sibling FE	Twin FE
Variable	coefficient / SE	coefficient / SE
Completed high school	0.2220*** (0.0070)	0.1437*** (0.0353)
Intercept	0.0617 (0.1383)	-0.1801** (0.0872)
Individual covariates	YES	YES
Number of observations	41,174	1,964
R2	0.0930	0.0358
IV	NO	NO

note: *** p<0.001, ** p<0.01, * p<0.05

From the table we find that both the sibling FE and the twin FE are lower than the LPM estimate from table 4 (26 percent). Hence, we would expect some of the estimated effect in the LPM estimate to be omitted variable bias alleviated by our sibling and twin fixed effects designs.¹⁵

In sum, we find very large effects for compliers using distance to nearest high school as IV for high school on the effect of high school on the propensity of move out of a rural area. Hence, individuals on the extensive margin are highly affected from getting a high school degree on their geographically mobility presumably because they have very low out migration rates without a degree and hence their optional value of a degree is very large. On average we find that outmigration rates go up with approximately 15 to 20 percentage points when one obtains a degree.

9. Conclusion

In this paper we have studied the effect of obtaining a high school degree on a population of adolescents living in rural areas on the propensity to move to an urban area. The

¹⁵ Some controversy roams in the economics literature on the external validity of twin studies, see footnote 12 and the associated main text. We have estimated random effects models on the twin and sibling data yielding much the same estimate as the original LPM estimate. Hence, from this, there is no evidence that the sibling and twin's data behave different from the total data set.

analysis is motivated by an increasing rural urban divide whereby the distribution of human capital is more and more skewed in favour of the urban areas.

It appears that one essential and very significant factor behind this divide is a large and increasing net migration of youth from rural to urban areas. Using Danish register data, we address the importance of education on the mobility between rural and urban areas. More specifically we address the causal effect of obtaining a high school degree on the probability of moving out of a rural area. Using distance to high school at the age of 15 as instrumental variable we find very large complier effects in the range of 65 – 85 percentage points. These effects are especially large for low SES students and wear off for higher SES groups.

The reason for the large complier effects is rooted in the fact that compliers – student that choose high school because of proximity – have very low a priori probability of moving and very high optional value of having a degree. These individuals appear to have their optimal choice set changed when obtaining a high school degree. Without a high school degree they might be more likely to stay unskilled or move into vocational education – both choices that makes staying in a rural area relative more attractive than when completing a high school degree, as the latter opens access to tertiary educational opportunities. These are mostly located in urban areas.

Our study provides strong evidence for the mobility effect of high school out of rural areas. We use a novel instrument, distance to nearest high school. Distance to educational institutions has been used in previous studies on the economics return to education but never in geographically mobility studies and never of the effects of a high school degree. Further, we refine our instrument in the light of potential validity problems. To overcome potential problems with distance to high school being endogenous – parents with high educational aspirations for their children locate nearer to high schools - we use only changes in distance to high school within families – that is the difference in distance to high school that occur between siblings when a high school institution open or close and thus change the distance to the nearest high school. If parents cannot forecast the openings and closure of high schools, the change in distance between siblings is arguably an exogenous chock. We believe we are the first to use openings and closures of educational institutions as an instrumental variable for educational choice. Although our change in distance is a much less efficient instrument it yields effect sizes in the same range as our original – and potentially more endogenous – instrumental variable.

Appendix

Appendix: Net migration rate

The net migration rate is calculated using the formula below:

$$N = \frac{I - E}{M} * 100$$

Where N denotes Net migration rate in percentage points. I denotes the number of immigrants (young people moving in) and E denotes the number of emigrants (young people between 18 – 24 years moving out). M denotes mid year population of young people (18 – 24 years), that is:

$$M = \frac{\text{Population at start of year} + \text{Population at end of year}}{2}$$

Appendix Tables

Table A1: Distance to nearest high school in 5 kilometers intervals

Distance in kilometres	Frequency	Percent	Cum. Percent
5	34,584	23.63	23.63
10	38,715	26.45	50.08
15	36,799	25.14	75.22
20	23,097	15.78	91.00
25	9,192	6.28	97.28
30	2,568	1.75	99.03
35	491	0.34	99.37
40	529	0.36	99.73
45	389	0.27	99.99
50	11	0.01	100.00
Total	146,375	100.00	

In table A2 we show LPM results for our four parental SES groups.

Table A2: Results of linear regression (LPM) by parent's education

Dependent variable: Moving to a large city before turning 25				
	LPM: SES1	LPM: SES2	LPM: SES3	LPM: SES4
Variable	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE
Completed high school	0.2609*** (0.0084)	0.2664*** (0.0040)	0.2535*** (0.0044)	0.2066*** (0.0081)
Intercept	0.0617 (0.1383)	-0.1863** (0.0818)	-0.1801** (0.0872)	0.2526** (0.1092)
Number of observations	21,084	71,141	43,302	10,848
R2	0.1774	0.1978	0.1917	0.1542
Individual controls	YES	YES	YES	YES
Family control	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES

note: *** p<0.001, ** p<0.01, * p<0.05

Robustness with regard to definition of migration:

We make robustness checks to our definition of urban area, see table A3 below. We run three robustness checks: One, where we redefine an urban area as one of the four largest cities in Denmark (Copenhagen, Aarhus, Aalborg and Odense), second, we define an urban area as cities with at least one university. Lastly, we rerun the analysis where we only look at moves to the Capital area, Copenhagen. The first two definitions of urban areas, do not change the results substantially.

Table A3: Robustness with regard to definition of dependent variable Y: inter-regional migration – LPM regressions

Robustness with regard to definition of dependent variable Y: inter-regional migration					
	Y1	Y2	Y3	Y4	Y5
	Move to large city	Move to one of the four largest cities	Move to university city	Move to Copenhagen city	Move to a large city in another municipality
Variable	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE
Completed high school at age 25	0.2641*** (0.0027)	0.3163*** (0.0028)	0.3090*** (0.0028)	0.1439*** (0.0025)	0.2863*** (0.0029)
Intercept	-0.1603*** (0.0512)	-0.6727*** (0.0528)	-0.5285*** (0.0533)	-0.0804* (0.0461)	-0.3392*** (0.0545)
Individual controls	YES	YES	YES	YES	YES
Family controls	YES	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES	YES
IV					
Number of observations	146,375	146,375	146,375	146,375	146,375
R2	0.2341	0.3243	0.2889	0.2699	0.2554

note: *** p<0.001, ** p<0.01, * p<0.05

Table A4: Robustness with regard to definition of dependent variable Y: inter-regional migration – IV regressions

Robustness with regard to definition of dependent variable Y: inter-regional migration, with IV					
	Y1	Y2	Y3	Y4	Y5
	Move to large city	Move to one of the four largest cities	Move to university city	Move to Copenhagen city	Move to a large city in another municipality
Variable	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE	coefficient / SE
Completed high school at age 25	0.6587*** (0.0750)	0.8784*** (0.0817)	0.7054*** (0.0777)	0.7085*** (0.0737)	0.2180*** (0.0749)
Intercept	0.3132*** (0.1052)	0.0019 (0.1146)	-0.0527 (0.1091)	0.5971*** (0.1035)	-0.4212*** (0.1051)
Individual controls	YES	YES	YES	YES	YES
Family controls	YES	YES	YES	YES	YES
Municipality FE	YES	YES	YES	YES	YES
IV	YES	YES	YES	YES	YES
Number of observations	146,375	146,375	146,375	146,375	146,375
R2	0.1245	0.1401	0.1941	0.0068	0.2525

note: *** p<0.001, ** p<0.01, * p<0.05

Table A5: Share of young adult that move to a large city before turning 25, split by education

		Stayed	Moved	Total
Vocational or no high school education	Freq.	33.225	33.398	66.623
	Percent	50%	50%	100%
High school education	Freq.	9.587	70.165	79.752
	Percent	12%	88%	100%
Total	Freq.	42.812	103.563	146.375
	Percent	29%	71%	100%

Table A6: Comparison of OLS and LOGIT estimation results

Dependent variable: Moving to a large city, before turning 25		
	LPM	LOGIT (AME)
Variable	coefficient / SE	coefficient / SE
Completed high school at age 25	0.2641*** (0.0027)	0.2239*** (0.0024)
Number of observations	146,375	146,375
R2	0.2341	0.2137
Individual controls	YES	YES
Family control	YES	YES
Municipality FE	YES	YES
IV	NO	NO

note: *** p<0.001, ** p<0.01, * p<0.05

AME: Average Marginal Effects

Figure A1: Research design

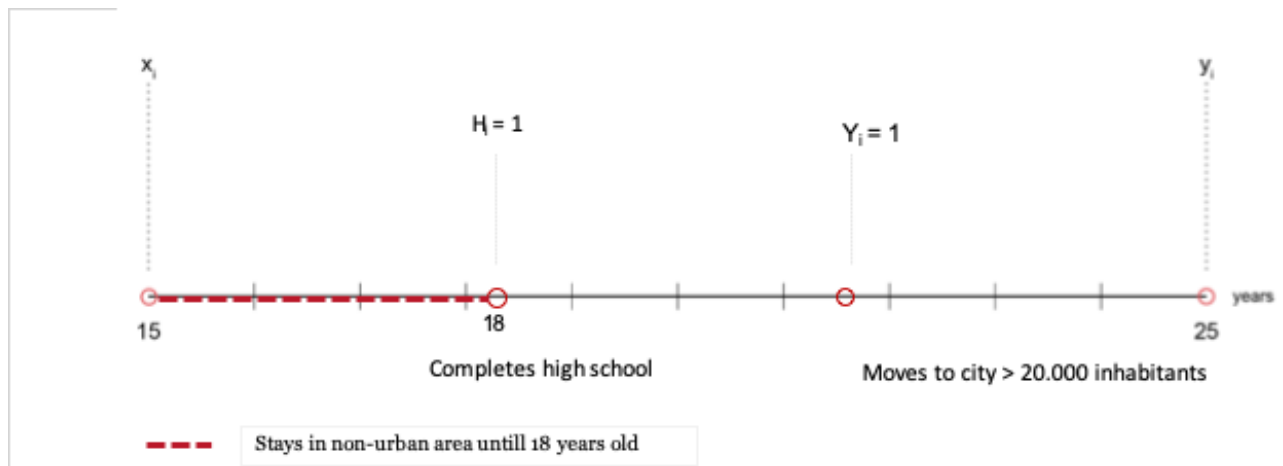
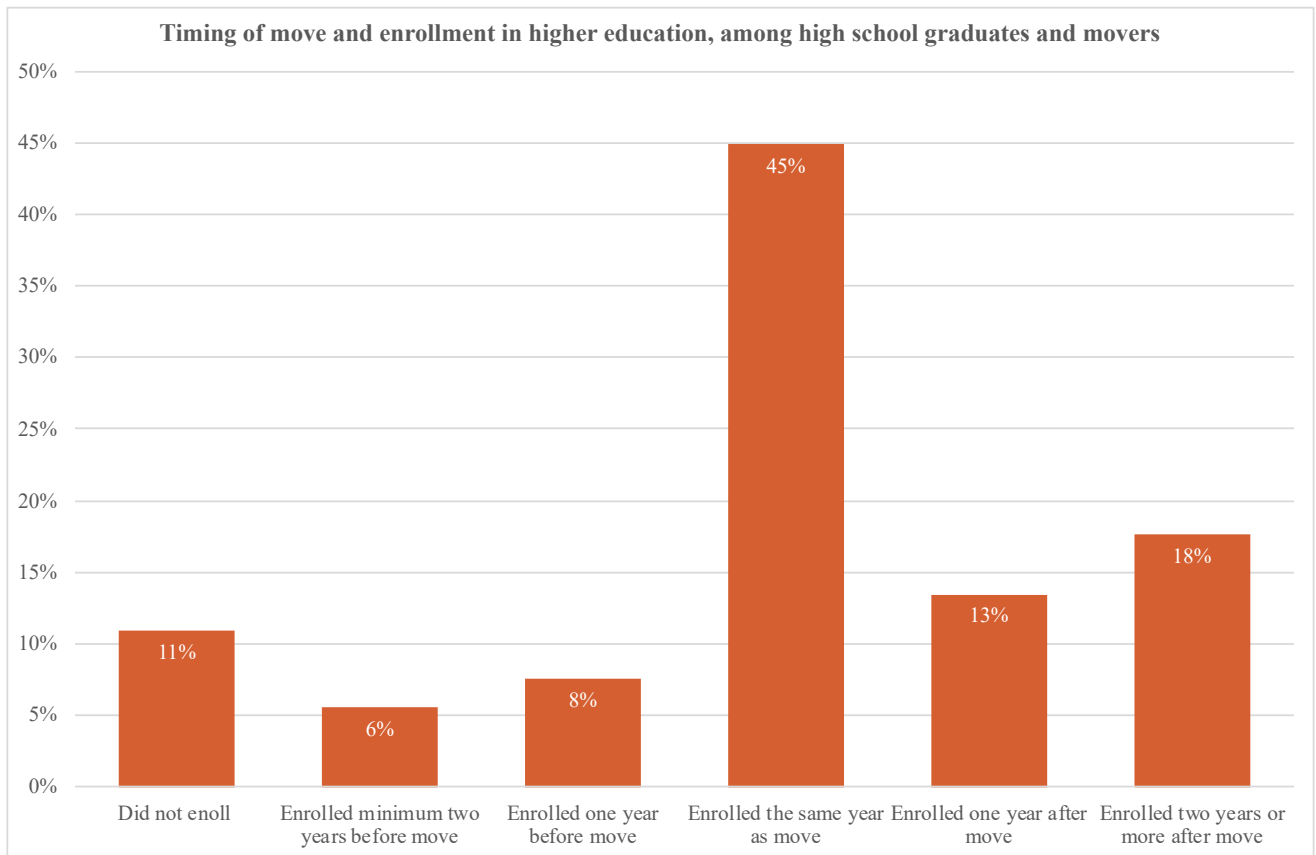


Figure A2: Correlation between moving to large city and enrollment in higher education. Sub-sample of high school graduates that moves (N = 70,165).



References

- Berck, P., Tano, S., & Westerlund, O. (2016). Regional Sorting of Human Capital: The Choice of Location among Young Adults in Sweden. *Regional Studies*, 50(5), 757–770.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430), 443–450.
<https://doi.org/10.2307/2291055>
- Busch, O., & Weigert, B. (2010). Where have all the graduates gone? Internal cross-state migration of graduates in Germany 1984-2004. *Annal of Regional Science*, 44, 559–572.
- Cameron, S. V., & Taber, C. (2004). Estimation of Educational Borrowing Constraints Using Returns to Schooling. *Journal of Political Economy*, 112(1), 132–182.
- Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, 201–222.
- Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating Marginal Returns to Education. *The American Economic Review*, 101(6), 2754–2781.
<http://dx.doi.org.ep.fjernadgang.kb.dk/10.1257/aer.101.6.2754>
- Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., & Wallace, B. (2009). Genetic Variation in Preferences for Giving and Risk Taking. *Quarterly Journal of Economics*, 124(2), 809–842.
- 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.

- Détang-Dessendre, C., Goffette-Nagot, F., & Piguet, V. (2008). Life Cycle and Migration to Urban and Rural Areas: Estimation of a Mixed Logit Model on French Data*. *Journal of Regional Science*, 48(4), 789–824. <https://doi.org/10.1111/j.1467-9787.2008.00571.x>
- Goldberger, A. S. (1979). Heritability. *Economica*, 46(184), 327–347.
- Haapanen, M., & Tervo, H. (2012). Migration of the Highly Educated: Evidence from Residence Spells of University Graduates. *Journal of Regional Science*, 52(4), 587–605. <https://doi.org/10.1111/j.1467-9787.2011.00745.x>
- Heckman, J. J. (2007). The economics, technology, and neuroscience of human capability formation. *Proceedings of the National Academy of Sciences*, 104(33), 13250–13255. <https://doi.org/10.1073/pnas.0701362104>
- Iammarino, S., Rodriguez-Pose, A., & Storper, M. (2018). Regional inequality in Europe: Evidence, theory and policy implications. *Journal of Economic Geography*. <https://doi.org/10.1093/jeg/lby021>
- Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2), 467–475. <https://doi.org/10.2307/2951620>
- Kodrzycki, Y. K. (2001). Migration of recent college graduates: Evidence from the National Longitudinal Survey of Youth. *New England Economic Review*, 13-+.
- Kotzeva, M. (Series Ed.). (2016). *Urban Europe. Statistics on cities, towns and suburbs, 2016 edition*. Luxembourg: Eurostat, European Union.
- Mellander, C., Florida, R., & Stolarick, K. (2011). Here to Stay—The Effects of Community Satisfaction on the Decision to Stay. *Spatial Economic Analysis*, 6(1), 5–24.
- Moretti, E. (2013). *The New Geography of Jobs* (Reprint edition). Boston, Mass.: Mariner Books.
- Pew Research Center. (2018). *What Unites and Divides Urban, Suburban and Rural Communities*. Washington DC.

- Rosenthal, S. S., & Strange, W. C. (2008). The attenuation of human capital spillovers. *Journal of Urban Economics*, 64(2), 373–389. <https://doi.org/10.1016/j.jue.2008.02.006>
- Royer, H. (2009). Separated at Girth: US Twin Estimates of the Effects of Birth Weight. *American Economic Journal: Applied Economics*, 1(1), 49–85. <https://doi.org/10.1257/app.1.1.49>
- United Nations. (2015). *World Urbanization Prospects: The 2014 Revision*. Retrieved from United Nations, Department of Economic and Social Affairs, Population Division website: <https://esa.un.org/unpd/wup/Publications/Files/WUP2014-Report.pdf>
- United Nations. (2018). *World Urbanization Prospects: The 2018 Revision*. Retrieved from United Nations, Department of Economic and Social Affairs, Population Division website: <https://esa.un.org/unpd/wup/Publications>
- United States Census Bureau. (2016, December 8). Our Changing Landscape. Retrieved July 12, 2019, from <https://www.census.gov/library/visualizations/2016/comm/acs-rural-urban.html>
- Van Der Gaag, N., & Van Wissen, L. (2008). Economic Determinants of Internal Migration Rates: A Comparison Across Five European Countries. *Tijdschrift Voor Economische En Sociale Geografie*, 99(2), 209–222. <https://doi.org/10.1111/j.1467-9663.2008.00454.x>
- Wooldridge, J. M. (2005). Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models. *The Review of Economics and Statistics*, 87(2), 385–390. Retrieved from JSTOR.