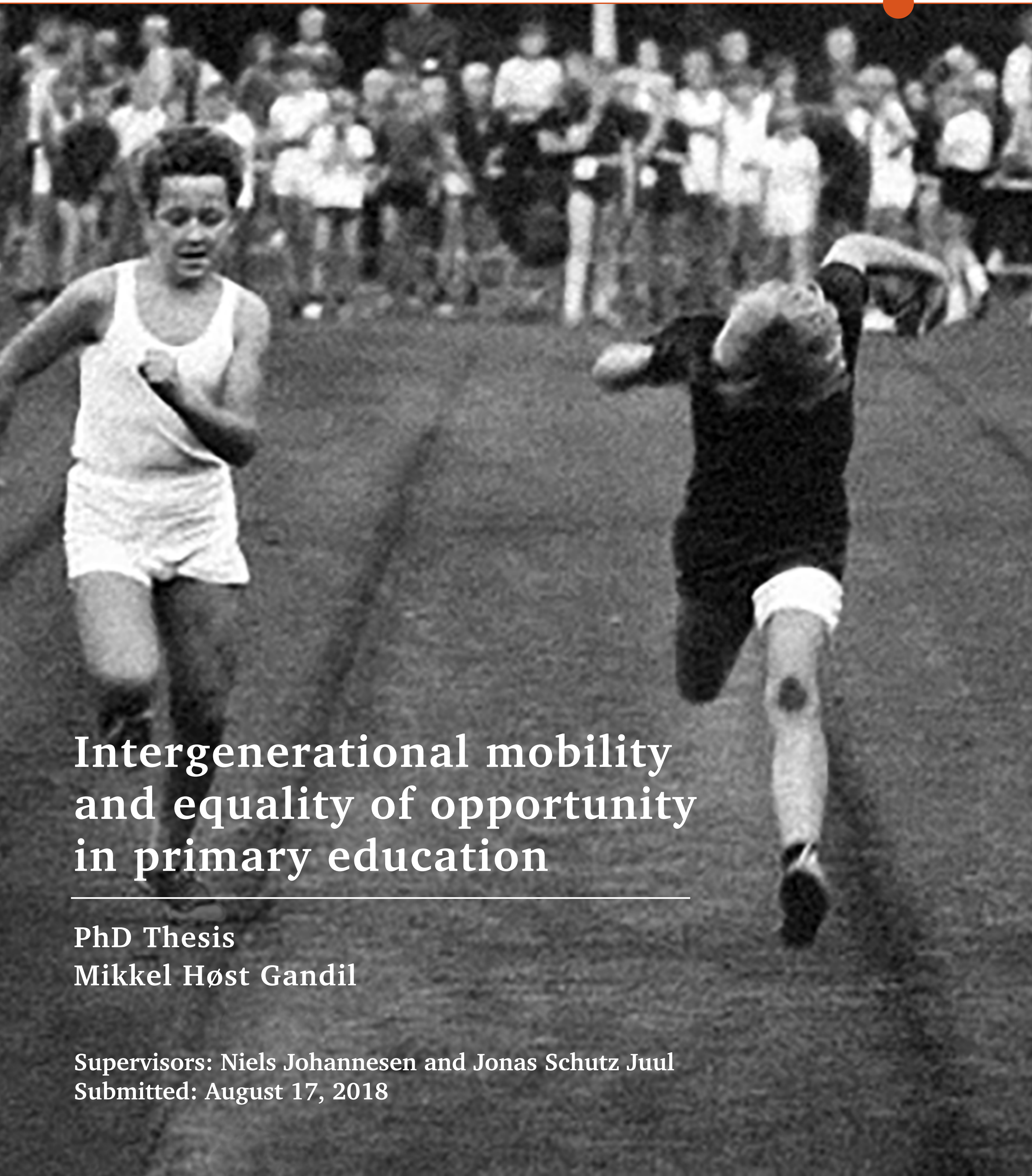
UNIVERSITY OF COPENHAGEN
FACULTY OF SOCIAL SCIENCES



Intergenerational mobility and equality of opportunity in primary education

PhD Thesis
Mikkel Høst Gandil

Supervisors: Niels Johannesen and Jonas Schutz Juul
Submitted: August 17, 2018

Contents

Acknowledgements

My road to becoming an economist has been, if not windy, then long. From being a somewhat old political science dropout to becoming an econ PhD, I have had the pleasure of encountering amazing people and brilliant minds, two sets where the union is barely larger than the intersection.

I would like to thank my university supervisor, Niels Johannesen, for teaching me the rigor by which an argument of casualty should be made and for his excellent ability to see the faults of an analysis and the possible remedies. I also want to thank my company supervisor, Jonas Schytz Juul, for his continued insistence on the importance of applicability and relevance of research for society.

As evident from the authorship of my chapters, this thesis would look much different had I not shared an office with Andreas Bjerre-Nielsen in the first year of my PhD. Besides being a skilled and open-minded economist with a work ethic out of this world, Andreas has become a good friend.

Doing my under-graduate studies my primary reason for not applying for a PhD was the thought of being lonely. I would like to thank myself for not realizing this mistake and applying sooner. In that case, I would not have had the opportunity to share offices (or semiprivate cubicles) at the Department of Economics with Adam Sheridan, Benjamin Ly Serena, Katrine Tofthøj Jakobsen, Louise Willerslev-Olsen and Sarah Clifford. I hope these people will continue to be the amazing sparring partners and good friends that they have been for the last three years. I would also like to thank Bjørn Meyer and Anders Priergaard Nielsen, the sharp wits of whom never seem to be affected by beer to the same extent as my own.

For inviting me to the Paris School of Economics and for great discussions I want to thank Daniel Waldenström. The stay at Jourdan was an extremely gratifying intellectual experience, not least due to the great people of office 313: Paul Dutronc, Anthony Lepinteur, Fanny Landaud and Simon Briole.

My employment in Arbejderbevægelsens Erhvervsråd frames my journey from BA to PhD. As a place of friendship, intellectual curiosity and professional development, AE has no rival in my adult life. Maybe most important of all, AE has continuously provided lessons in the potential for economics to be wielded for the betterment of the common good and fairness in society.

I would like to thank my parents and my grandmother who throughout my life have always provided love and support. I would not have become a social scientist had it not been for countless engaging, open-minded, and (in the very best connotations of the word) challenging discussions that they have always been willing to engage in.

Lastly, I would like to thank Ida whose kindness and brilliance never seize to amaze me.

Mikkel Høst Gandil
Copenhagen, August 2018

Dansk introduktion

Denne afhandling handler om, hvordan økonomisk ulighed går i arv fra forælder til barn. Man kan ikke bebrejde et barn for sit valg af skole, materielt afsavn eller sin adgang til rollemodeller. Ethvert samfund må derfor interessere sig for, hvordan social position overføres fra generation til generation. Hvordan sker dette i praksis, er det fair og i modsat fald hvad kan gøres for at ændre det? Det er de spørgsmål, læseren bør have på sinde.

Afhandlingen består af to dele med fire artikler i alt. Fra forskellige vinkler omhandler alle artiklerne spørgsmål om lighed i muligheder og social arv. Først en enkelt artikel om målingen af social mobilitet, derefter tre kapitler, der beskriver, hvordan grundskolens indretning og opdeling af boligområder påvirker samfundets muligheder for at skabe lighed i adgangen til uddannelse og dermed lighed i muligheder.

Første del: Sammenhængen mellem mobilitet og ulighed.

Kapitel 1 er et metodebidrag om, hvordan mobilitet påvirkes af kønsulighed, sådan at det er svært at sammenligne samfund på tværs af tid og rum. Nogle gange fremkommer statistiske sammenhænge per konstruktion, og i de tilfælde bibringer sammenhængene ikke substantiel viden. I kapitlet viser jeg, hvordan et specifikt mål, kaldet rank-mobiliteten, stiger og falder sammen med indkomstuligheden mellem kønnene. Implikationen er, at et samfund, der diskriminerer kvinder, vil fremstå mere mobilt end et samfund, hvor kønsuligheden er blevet mindsket. Jeg viser, at mobiliteten i USA i den sidste halvdel af det tyvende århundrede skulle være faldet med 25 procent alene på grund af den større lighed mellem kønnene. Kønsdiskrimination vil dermed fremstå positivt, idet det mindsker betydningen af social arv. Hvis dette synes absurd, bør resultaterne i dette kapitel mane til forsigtighed med at bruge rank-mobiliteten til at måle og sammenligne social arv.

Anden del: Lighed i muligheder, når forældrene vælger skolen.

I halvfemserne konstruerede en række økonomer modeller for, hvordan opsplitningen af boligområder kan påvirke den sociale arv. Det centrale i disse modeller er, at mennesket ikke skal ses i isolation fra dets omgivelser men derimod som en del af fællesskaber i boligområder, arbejdspladser og skoler. Disse sociale og økonomiske fællesskaber er centrale for forældres muligheder for at give deres sociale status videre til næste generation. For en socialvidenskaber vil dette ikke være en overraskende indsigt, men Bénabou (1993), Durlauf (1996a,b) og Epple and Romano (1998) viste, at eksistensen af disse medlemskaber er vigtig for, hvordan man bør tænke på ulighed, mobilitet og muligheder for omfordeling.

En kort fremstilling af argumentet lyder som følger: Hvis samfundet kunne sikre alle elementer, der indgår i et barns uddannelse, ved at betale for dem, ville traditionel omfordeling mellem skoler kunne sikre lige adgang til uddannelse. Men hvis børn påvirker hinanden i skoler, er det ikke længere så simpelt. Hvis stærke børn trækker svage børn op, er det pludselig vigtigt hvilken skole, børnene går på. Fordi skoler i bund og grund er lokalt definerede institutioner, repræsenterer geografien en grundlæggende forudsætning for at ulighed kan gå i arv. Hvis stærke og svage børn ikke går på de samme skoler, kan udsatte børn ikke nyde godt af den positive påvirkning fra de stærke børn. Dermed har de udsatte børn ikke mulighed for at opnå deres fulde potentiale. Dette uagtet at samfundet måske gør sit bedste for at udligne økonomiske forskelle i stærke og svage skoler.

De tre artikler i denne del er skrevet sammen med Andreas Bjerre-Nielsen og undersøger, om dette er en empirisk relevant problemstilling i Danmark. Det overordnede spørgsmål er dermed, hvordan ulige boligområder og skoler påvirker et af verdens mest veludviklede velfærdssamfunds muligheder for at sikre lige adgang til uddannelse for alle.

Hvis børn påvirker hinanden i klasseværelset, bør det lede beslutningstagere til at prøve at påvirke elevsammensætningen på skolerne. Ved at påvirke elevsammensætningen kan man muliggøre, at børn fra forskellige kår mødes og indgår i de samme sociale fællesskaber. En måde kommuner gør dette i praksis er ved hjælp af skoledistrikter. I kapitel 2 viser vi, at det dog ikke er så nemt i praksis. Ressourcestærke familier modsætter sig kommuners forsøg på at ændre skoletilknytning, når dette medfører, at deres børn skal gå i klasse med mere udsatte børn. Til vores overraskelse er privatskoler ikke særlig vigtige i denne sammenhæng. Ressourcestærke forældre udnytter i stedet det frie skolevalg til at undgå distriktsskolen. Ressourcesvage forældre tager derimod mindre hensyn til elevsammensætningen, når de vælger, om de vil udnytte muligheden for frit skolevalg. Vi kan ikke med sikkerhed fastslå, hvorfor dette er tilfældet, men har dog en kraftig mistanke om, at en af grundene er, at frit skolevalg er et uigennemsigtig og svært system at navigere i. Det gør, at ressourcestærke familier har nemmere ved at få deres ønsker til skole opfyldt, end ressourcesvage har, og resultaterne bør lede til overvejelser om, hvorvidt frit skolevalg kan gøres mere gennemsigtigt og fair.

Skulle man vælge at ændre systemet med skoledistrikter og frit skolevalg, er det dog ikke nødvendigvis nok til at sikre lige adgang til uddannelse. Selvom folkeskoler er gratis i Danmark, kan forældre stadig betale sig adgang til skoler via ejendomsmarkedet. I den internationale forskning er det bredt anerkendt, at boligpriserne afhænger af den lokale skole. I kapitel 3 viser vi, at dette også er tilfældet i Danmark: des stærkere socioøkonomisk baggrund blandt eleverne i distriktsskolen, desto højere boligpriser. Vi isolerer denne effekt ved at bruge detaljeret geografisk data og undersøger, hvordan to ellers ens boliger varierer i pris som følge af forskelle i skoletilknytning. Derefter bruger vi ændringer i distrikterne til at vise, at priserne tilpasser sig en ny skoletilknytning inden for tre år. Med disse resultater kan vi beregne, at ressourcesvage familier må afgive en substantiel andel af deres indkomst for at købe sig adgang til ressourcestærke skoler. Selv i en velfærdsstat med offentlig finansiering af grundskolen skaber geografien derfor afgørende hindringer for lige adgang til uddannelse og stærke klassekammerater.

Er ressourcestærke familier så rationelle, når de vælger socioøkonomisk svage skoler fra? Dette er emnet for kapitel 4. Ændringer i skoledistrikter gør, at børn potentielt udsættes for andre klassekammerater, end deres forældre oprindeligt tiltænkte, da de valgte, hvor de skulle bosætte sig. Dette bruger vi til at undersøge, hvordan børn klarer sig, afhængigt af hvilken baggrund deres klassekammerater har. Overordnet finder vi, at børn ikke påvirkes i nævneværdig grad. Der er dog én gruppe, hvor dette ikke er tilfældet: de udsatte børn, som vi ville forvente ville benytte distriktsskolen i alle tilfælde. Denne gruppe har gavn af at komme på skoler med flere ressourcestærke elever. For de ressourcestærke børn dokumenterer vi nul effekt. Med andre ord finder vi ingen negativ påvirkning fra de svage elever på de stærke elever. Der er mange måder, hvorved de positive effekter for svage elever kan opstå. Børn kan påvirke hinanden i klasseværelset og frikvarteret, men forskellige forældre stiller også forskellige krav til lærere og faciliteter. Vi kan med vores data ikke sige præcist, hvad der skaber denne positive effekt, men vi finder kraftige indikationer på, at den eksisterer. Givet at ressourcestærke ikke påvirkes af deres klassesammensætning, giver det mulighed for, at man ved at blande elever fra forskellige kår kan forbedre udsatte børns muligheder. Vel og mærke uden at andre børn lider skade.

Hvad kan man da konkludere, efter at have læst de tre artikler i denne del? I det sidste kapitel har vi vist, at klassekammeratseffekter højst sandsynligt virker på en måde, som gør at ulige skoler skaber ulighed i muligheder og dermed i livet. Samtidig har vi vist, at der er potentiale for at forbedre mulighederne for alle ved at omfordele elever på skoler. Men kapitlerne viser også, at der er klare forhindringer for at gøre dette i praksis. Forældre har holdninger til, hvilken skole deres børn skal gå på, og de har muligheden for at modarbejde omfordelingen af skolebørn dels via frit skolevalg og privatskoler og dels via ejendomsmarkedet. De to førstnævnte kanaler kan måske ændres politisk, men som altid vil det være svært at sikre at forældres mulighed for at udnytte systemer ikke vil reproducere social ulighed. Skulle det lykkes at implementere fair systemer, vil der imidlertid være en grænse for, hvor langt fra hjemmet børn kan gå i skole. Uden en aktiv boligpolitik, vil ejendomspriserne derfor altid være en hindring for at skabe lighed mellem skoler. Denne afhandling viser dermed, at selv i en af verdens måske mest veludviklede velfærdsstater er det mere end svært at sikre lige muligheder for alle børn, uanset baggrund.

Alle fire kapitler kan læses uafhængigt af hinanden og efterfølges af litteraturliste og bilag. De tre kapitler om skoler er baseret på nøjagtig geografisk data om boliger og individer. For at kunne få adgang til dette data, har Andreas og jeg udviklet en algoritme til at sikre anonymitet. Denne algoritme er dokumenteret i et teknisk bilag.

English introduction

This thesis is concerned with a possible answer to the question of why we should care about inequality: children. A child can no less choose its parents than it can choose its gender or ethnicity. One cannot fault a child for their choice of school, lack of resources or absence of role models. Therefore, any meritocratic liberal society must grapple with how inequalities persist across generations. In other words, how do parents transmit social status to children? To what extent is such transmission fair, or not, and what can society do to change it? While reading this thesis, these are the questions to have in mind.

This thesis is divided into two parts, with four articles in total which, all relate to the question of equality of opportunity and social mobility. The first chapter is self-contained and focuses on the statistical measurement of intergenerational mobility while the remaining three chapters focus on the importance of primary education, inequality in educational input in childhood and its consequences.

Part 1: The link between measurements of inequality and mobility. Statistical associations sometimes prove to be a product of methodological definitions where subtle assumptions create links in a way that may provide little real insight. Chapter 1 is a methodological contribution. I investigate how positional mobility in the income distribution and gender inequality is mechanically linked in a way that makes it hard to compare mobility across time and across space. Using a popular measure, the rank correlation, I show that mobility rises mechanically with gender inequality. The implication of this mechanical link is that a society, which discriminates against women, may appear more mobile, and thus more desirable, than a society which has succeeded in diminishing inequality.

Using real income distributions, I find that this finding matters in practice. For a constant parent-son and parent-daughter rank-mobility, the diminishing gender inequality in the last 40 years would have diminished mobility by 25 percent. In the chapter, I argue that these findings warrant careful reconsideration of the rank-correlation as a mobility measure.

Part 2: Equality of opportunity when students sort. The second part of the thesis present three closely related papers. These are empirical papers, but this introduction provides a possibility to tie down how they relate to the overarching theme of inequality and mobility. In the nineties, a group of economists constructed models of how neighborhood formation can be a prime conduit of inequality from one generation to the next. The key component of these models is that, far from being free-floating agents, humans are parts of social groups, workplaces, neighborhoods and schools. This observation may not be surprising for the common observer, but Bénabou (1993), Durlauf (1996a,b) and Epple and Romano (1998) showed that such group memberships have great implications for how to think about inequality, social mobility and redistribution.

If child outcomes depended solely on educational inputs, which could feasibly be financed by the government, then (assuming political will) traditional redistribution between schools may be sufficient to achieve full equality in access to education. However, insofar as strong students pull up weaker students, this is no longer the case and sorting of children into schools becomes important. Disadvantaged children can be shut off from potential benefits if they are not enrolled in schools with strong peers. This, in turn, may lead to inequality in human capital and earnings. Inequality can thereby transmit from the parent to the child generation through the exclusivity of schools.

The three chapters in the second part, written together with Andreas Bjerre-Nielsen, all center on how these dynamics of sorting and social interactions may cause inequality in access to primary schools in Denmark. Danish children are allocated to schools via attendance boundaries. For an empirically inclined economist, this provides a range of natural experiments, especially as administrators, surely in their benevolence towards my professional caste, choose to change these boundaries over time. This is the variation which forms the basis of the empirical analysis in the three chapters.

From a policy perspective, any effect of student interaction naturally leads to a consideration of the scope for affecting sorting into schools, i.e. affecting where and how children interact. The first chapter in this section, chapter 2, investigates the scope for such policies in Denmark, namely changing the socioeconomic compositions of student bodies by way of manipulating attendance boundaries. We document that households respond in ways that negate intended effects of the boundary changes. The results indicate that especially households of high socioeconomic status do not comply with the reassignment when the child would potentially be exposed to disadvantaged peers. A municipality is therefore constrained in its capabilities to heighten the socioeconomic status of a given school. Contrary to our expectations, we show the private schools, though abundant in Denmark, play a minor role in explaining the behavioral responses. Instead, a publicly provided loophole explains most of the behavioral response from the families with strong socioeconomic backgrounds. Though we cannot be sure from the data, we conjecture that the lack of responses from disadvantaged households stem partly from preferences and partly from the inability to navigate the bureaucratic procedures needed to exploit the loophole. Thus, the source of inefficacy stems directly from previous policy choices, which may be amended in the future.

The behavior of households in chapter 2 shows that parents have preferences over schools. But if so, they should be willing to pay for schools. Parental co-payment of public schools is forbidden by law in Denmark, but parents still pay for schools through the housing market. In the international literature, it is an established finding that house prices reflect the quality of the local schools. If this is the case, then even in the case where public education is provided free of charge, the interaction between income inequality and housing markets will create unequal access to education. In chapter 3 we show that a higher socioeconomic status of schools is associated with higher housing prices. To isolate these premiums, we use highly detailed geographical data and changes in school attendance boundaries over time. We show that sales prices adjust to a new school association within three years. By back-of-the-envelope calculations, we find that poor households must give up a significant fraction of their disposable income to ensure enrollment in a school with strong socioeconomic peers. The take-aways from this chapter is that even in a highly egalitarian society such as Denmark, access to education is not equal; high-income households can buy houses close to desirable schools and thereby provide their children with stronger peers to the detriment of children from low-income households.

Are privileged families behaving rationally when they exploit the institutions and the housing market to avoid schools to which disadvantaged families may send their children? This is the subject of the chapter 4. Regardless of behavioral responses to reassignments documented in chapter 2, most children still enroll in the school intended by authorities. We, therefore, exploit the changes in attendance boundaries to investigate the importance of exposure to peers different from those that parents originally intended. Measured by the performance on low-stake language tests, the analysis shows that children *in general* are not affected by the socioeconomic composition of their peers. We find that children from educated and affluent families are especially *insensitive* to intended changes in the peer composition of the local school. However, for a small subset of children, we document strong effects. Children from disadvantaged families benefit from attending schools with children from advantaged backgrounds, but only if they are likely to enroll in the local school in the first place. Why these disadvantaged kids gain, we cannot say. Exposure to stronger peers can affect children both by direct interaction and through parental demands for educational services. Regardless of the mechanism, the implication is that outcomes of disadvantaged children can be improved at little or no cost to other children. In other words, besides considerations of fairness, overall economic efficiency might be improved by equalizing student compositions across schools.

So, what is the overall conclusion of these three chapters on Danish primary schools? There are strong indications that interaction between children matter in a way such that segregation creates both unequal opportunity and economic inefficiency. It is therefore likely that there is no trade-off between equality and efficiency. By decreasing sorting in schools, policymakers can, therefore, increase economic mobility and grow the economy at the same time. However, parents negate the effects of such associational redistribution by exploiting markets and loopholes. These unintended consequences constitute severe restrictions on the means by which society can achieve equality of opportunity.

The four chapters in this thesis are self-contained with abstracts, bibliographies and appendices. The three chapters on Danish schools all make use of geocoded, individual data. Due to anonymity protection, Andreas and I have had to develop an algorithm to ensure anonymity in order to gain access to this data. The algorithm is documented in a technical appendix to the thesis.

Bibliography

Bénabou, R. 1993. Workings of a city: location, education, and production. The Quarterly Journal of Economics, 108, 619–652.

Durlauf, S. N. 1996a. A theory of persistent income inequality. Journal of Economic Growth, 1, 75–93.

Durlauf, S. N. 1996b. Associational Redistribution: A Defense. Politics & Society, 24, 391–410.

Epple, D., Romano, R. E. 1998. Competition between Private and Public Schools, Vouchers, and Peer-Group Effects. The American Economic Review, 88, 33–62.

Chapter 1

Intergenerational mobility or gender inequality: What are rank correlations measuring?

Intergenerational mobility or gender inequality: What are rank correlations measuring?

Mikkel Høst Gandil

Abstract

The rank correlation between parent and child has become a central measure of intergenerational income mobility in recent years. A great advantage of the rank correlation, compared to other measures such as elasticities, is the invariance with respect to the shape of income distributions and therefore intragenerational inequality. This facilitates comparisons of societies over time and across space. However, I show that changes in inequality between genders *within* generations directly affect the rank correlation; diminishing gender inequality leads to a fall in mobility. The implication is that the same rank correlation can map into very different societies. By estimating American income distributions and assuming constant within-gender mobility, I show that the rank-correlation should have risen by almost 25 percent over the last 40 years, solely due to the narrowing gender income gap. An arguably benign decrease in gender inequality will therefore register as an adverse development in measured mobility. The findings underscore the importance of being explicit about the definition of income concepts and the importance of group-inequality when comparing rank correlations between countries and across time.

1 Introduction

Economic inequality and intergenerational mobility has taken center stage in recent years.[1] This renewed public interest, as well as access to new data sources, have reinvigorated the inequality research agenda. The most popular measure of mobility is the intergenerational income elasticity (henceforth abbreviated IGE), which measures the link between parent and child income. Recent research has documented close links between inequality and this measure of mobility. The "Great Gatsby-curve" may be the best know example.[2]

However, the IGE is itself a function of inequality. There may, therefore, be mechanical linkages between inequality within generations and the IGE. If this is the case, interpretation of a difference over time or between societies becomes difficult. Furthermore, the IGE has been shown to be quite sensitive to the choice of income concepts and modeling choices, see Chetty et al. (2014b) for an example. To address this problem, part of the literature has shifted focus from the actual income to *positions* in the income distribution. By only comparing the intergenerational association between positions, also called ranks, and not actual incomes, one may disregard differences in the shape of income distributions. Rank-based mobility measures, therefore, facilitate comparisons of mobility across time and between countries while disregarding inequality. In other words, mobility and inequality become decoupled into two separate concepts. Such a measure of association between parent and child rank is the rank correlation, also known as the Spearman correlation coefficient.

Chetty et al. (2014a) show that the rank correlation in America has been remarkably stable for the last decades, despite sizable increases in cross-sectional inequality. These considerations could lead one to con-

[1]As then American president, Barack Obama, stated on December 4, 2013: *"The combined trends of increased inequality and decreasing mobility pose a fundamental threat to the American Dream, our way of life, and what we stand for around the globe."*

[2]Corak (2006) documented that countries with high cross-sectional inequality also tend to have lower intergenerational mobility. In other words, a high level of inequality is associated with a stronger dependence between parent and child income. The term "Great Gatsby curve" came later than the finding. It was first introduced in 2012 by Alan Krueger, then chairman of the Council of Economic Advisors.

clude that the rank-correlation is superior to the IGE as a measure of intergenerational mobility. However, I argue in this paper that the robustness of the rank-correlation may be overstated. I point to an often-overlooked issue of inequality between groups but within the same generation. When income distributions differ between groups, the interpretation of the correlation becomes ambiguous.

In this paper, I show that a society with high gender inequality will have higher rank mobility than a society with low gender inequality. This is the case even if the two countries exhibit the same relative mobility levels within gender. Intuitively, the importance of gender naturally depends on the level of gender inequality. If a society has no gender inequality, the gender of a child will matter little when we measure the association between the positions of parent and child in society. On the other hand, if a society has high gender inequality, then the sorting into gender becomes important. As the gender of a child is essentially random and uncorrelated with income, the gender gap will register as mobility from one generation to the next. This matters for comparisons over time and space. If a society manages to remove some structural obstacles for women in the labor market and this leads to less gender inequality, this will also lead to an apparent *fall* in mobility.

Formally, I exploit that rich and poor households alike share the randomness of a child's gender. I, therefore, conceptualize gender as a random sorting mechanism into two distinct groups; male and female. I then proceed to describe theoretically how the link between the mobility within the two genders and the mobility in society as a whole is a function of the level of gender inequality. I approximate the link as a linear relationship, where the factor is a function of the income distributions of men and women. Under reasonable assumptions on the shape of the income distributions of men and women, I show that this factor approaches one *from below* when the gender gap narrows. A given rank correlation can thereby map into infinitely many combinations of gender inequality and within-gender intergenerational mobility. Consequently, when gender is not taken into account we cannot know how to draw out policy implications from a change in the rank correlation.

To gauge whether this issue matters in practice, I take the results to

American survey data. Using the developed formulas, I show that the rank-correlation based on individual incomes should have risen by almost 25 percent due to the narrowing gender income gap alone. In other words, without any change in mobility *within* genders, the apparent mobility would have fallen substantially due to the increase in gender equality. As I use Taylor-approximations to develop the link between within-gender and aggregate mobility, there may be approximation errors. To assess this, I perform simulations of the link between within-gender mobility and aggregate mobility. I show that results from theoretically derived formulas and simulations align and that the overall conclusions are insensitive to the approach taken. The simulation tools I present in this paper may also serve as a tool for researchers to perform *ceteris paribus* analysis of the importance of inequality for explaining changes in mobility, and I make these tools freely available.[3]

As mentioned, the advantage of the rank correlation, as a measure of intergenerational mobility, is its supposed invariance to the shape of income distributions. My findings, however, show that while the aggregate income distribution may not matter, the shapes of the gender-specific income distributions certainly do. The issue highlighted in this paper directly concerns rank correlations when individual incomes are used and both genders are included in the same calculation. A way to bypass this problem is to estimate rank correlations for sons and daughters separately. Fortunately, this is often done when investigating changes over time. Chadwick et al. (2002) conjecture that the focus on sons might partly be due to unrealized sexism and the recognition that the entrance of women into the labor force present one of the most salient and fundamental developments of the latter half of the twentieth century.[4] For comparisons between countries, it might more have been due to luck than intent that the rank correlation is only calculated for sons. However, I have not found a thorough discussion on the implication of group in-

[3]Software to perform copula simulation is available from my website.

[4]A related feature may be the IGE's inability to handle zeros due to its' logarithmic nature, the sensitivity to year-to-year fluctuations and bias from mismeasurement of lifetime incomes, see Mitnik et al. (2015) and Haider and Solon (2006) for a discussion of some of these issues.

equality for the rank correlation.[5] An alternative strategy is to use household incomes and the unit of analysis. If household income is used one naturally does not observe the same degree of gender inequality. However, with rising trends in single adult households across the developed world and the substantial variation in household sizes across countries, gender inequality continues to pose problems for comparisons of rank correlations over time and space.

The issues presented in this paper directly ties into the discussion of how intergenerational mobility should be conceptualized. An extreme position is that gender inequality *actually is* mobility. The gender of a child is a lottery where the odds are the same for poor and rich households alike. As I show, this lottery decreases the importance of parental background characteristics. In other words, the randomness of gender reduces the importance of the accident of birth. However, the policy implication is that in order to make a society more mobile, gender inequality should be *increased*. By implication, rank-based measures cannot ignore gender inequality. The issues presented in this paper should, therefore, lead to thorough sensitivity analysis and methodological considerations of the choice of income concepts and unit of analysis when using rank-correlations in practice. In this way, the rank-correlation may not be easier to use than the traditional measures of mobility.

The paper proceeds as follows. Section 2 presents the canonical conceptual framework for studying income mobility along with measures of mobility with a focus on the rank correlation coefficient. I then introduce the issue of gender inequality into this framework. Section 3 shows how gender inequality affects the measure of mobility while section 4 quantifies the effect of gender inequality on intergenerational mobility. Section 5 discusses the generalization of the results while section 6 concludes.

[5]An example, where such considerations are not taken, is Landersø and Heckman (2017), where rank measures are employed for both genders jointly with individual incomes.

2 Measuring intergenerational mobility

I begin by introducing a standard framework in which to conceptualize inequality and intergenerational mobility. This framework follows Jäntti and Jenkins (2015) closely. Let a family be defined by a parent and a child income, x_i and y_i respectively. Most measures of intergenerational mobility can be thought of as describing the joint distribution of (x_i, y_i). Denote this bivariate joint distribution as $H(x, y)$ with the corresponding marginal distributions, $F(x)$ and $G(y)$. The marginal distributions are the income distributions of the two generations. Thus all measures of inequality, such as the Gini coefficient, can be calculated from F and G to describe inequality in the parent and child generation respectively.

Typical measures of mobility are based on a normalized covariance between parent and child income or log income. Pearson's linear correlation coefficient is defined as $\frac{Cov(x,y)}{\sigma_x \sigma_y}$ and the intergenerational income elasticity (IGE) is given by $\frac{Cov(\log x, \log y)}{Var(\log x)}$. The latter, to an approximation, describes the percentage change of child income when parent income is raised by one percent. Intuitively, the higher the IGE, the higher is the dependency between parent and child, and thus the lower is the intergenerational mobility. In practice, one usually obtains the IGE by regressing the logarithm of child income on the logarithm of parent income. The IGE has long been the preferred measure of income mobility, see Jäntti and Jenkins 2015 for an exhaustive review and Mitnik et al. (2015) for a recent application.

From the definitions of the correlation coefficient and the IGE, one can see that, if incomes are measured in logs, the two measures only differ by a rescaling.[6] The presence of the standard deviation of child income in the denominator of the correlation coefficient can be seen as a normalization, such that one may compare societies with different levels of inequality. Even so, Chetty et al. (2014b) make a convincing case for why these measures are difficult to interpret in practice. When using actual incomes as the unit of analysis, one cannot have a change in the

[6] To see this, let the Pearson correlation of log income be given as $\rho = \frac{Cov(log(x), log(y)}{\sigma_{log(y)}\sigma_{log(x)}} = \frac{\sigma_{log(x)}}{\sigma_{log(y)}} IGE.$

income distributions without directly affecting the joint distribution and thus measures of mobility based on income. An example could be general economic growth or rising inequality over time. This is a well-known property of mobility measures and has been shown to be empirically of great importance, see Jäntti and Lindahl (2012).

However, comparisons are not only made over time but across countries. Thus, differences in economic development and economic institutions between countries may directly enter into the calculation of the IGE. This property of the IGE is not necessarily a problem; the effect of institutions on mobility may truly be the object of interest in an analysis. Nevertheless, it is difficult to know whether the effect of institutions go through affecting inequality, i.e. the marginal distributions of $H(x, y)$ or through the dependence structure between parent and child.[7]

2.1 Disentangling mobility and inequality

In order to disentangle inequality and mobility, a branch of the mobility literature has moved towards a focus on the *position* in the income distributions, i.e. the ranks of the individuals. Denote the ranks of parent and child as u_i and v_i respectively. The ranks are most often calculated by software simply by sorting the data. Ranks can, however, be related to the marginal distributions by the *probability integral transform*; Assuming the income distributions are continuous, a rank is simply the cumulative distribution function applied to the income, $u_i = F(x_i)$ and $v_i = G(y_i)$. It follows that both u_i and v_i are always uniformly distributed. The rank is therefore invariant to monotonic positive transformations of the income distributions (i.e. changes in inequality) and changes general income lev-

[7]This point also relates to the interpretation of the Great Gatsby curve shoving a negative association between inequality and mobility measured by the IGE. Both measures are based on marginal distributions, and we may suspect a somewhat mechanical relationship. Berman (2017) investigates this point in setting with log-normal income distribution. He finds, that there may be such a relationship present. I have simulated the relationship between mobility and inequality with copulas and find that this relationship is somewhat mechanical. The sign of the slope, however, depends on the shape of the copula. Simulation results are available upon request.

els.[8]

The rank correlation, at times referred to as Spearman's ρ, is simply the linear correlation coefficient between parent and child rank;

$$\rho^S = \frac{Cov(u,v)}{Var(u)} = \frac{Cov(u,v)}{\sigma_u \sigma_v} \tag{1}$$

A link between the intergenerational dependence structure of incomes and ranks is provided by Sklar's theorem stating that any joint multivariate distribution can be described by the marginal distributions and a distribution describing the dependence of the ranks.[9] The latter distribution is called a copula and is defined on the unit square.[10] Formally $H(x,y) = C(F(x),G(y)) = C(u,v)$. It is, therefore, possible to apply rank-based measures of mobility while abstracting from the shape of income distributions. This facilitates comparisons of mobility over time and between countries regardless of differing income distributions. In other words; with the rank correlation, we can take inequality out of the picture and focus exclusively on mobility.

Besides rank correlations, other measures can also be derived from the copula. An example is transition matrices, which can be conceptualized as discretized representations of the copula. These types of measures also exploit the uniform marginals in order to make comparisons across societies. However, the rank correlation is the measure of interest for the rest of this paper.

2.2 The role of gender inequality

The framework described above is purely descriptive and is uninformative about mechanisms creating a given dependence structure. In this paper, I investigate one of these mechanisms, namely gender. Gender is a salient and well-defined characteristic and a gender wage gap is well doc-

[8]An issue concerns mass points in the income distributions such as zeros. I abstract from these in the present context, as I seek to establish a general point. How to treat zeroes is however massively important in applied research, see Chetty et al. (2014b) for an example.

[9]The theorem was introduced in Sklar (1959).

[10]For a proper proof see Nelsen (2006)

umented in most modern societies.[11] The gender gap in incomes has been greatly reduced since the middle of the twentieth century both through women entering into the labor force and a rise in relative education levels for women. In the following, I am agnostic as to what has caused this narrowing income gap and focus on the implications for measures of mobility.

In what follows, I assign the superscript m to families with a male child and f to families with a female child. Furthermore I assume that the gender of a child is independent of income, and that the gender is assigned by the random variable s_i, such that child i is a boy if $s_i = 1$.[12] Let $E[s] = \mu$, that is the share of males in a cohort. I assume this to be stable at 0.5. Gender assigns the child to an income distribution. Family i draws two sets of incomes where only one set is realized depending on gender: $(x_i, y_i) = s_i(x^m, y^m) + (1 - s_i)(x^f, y^f)$. With these assumptions, we can rewrite the joint distribution H:

$$H(x, y) = \mu H^m(x, y) + (1 - \mu)H^f(x, y)$$

and the marginal distributions F and G:

$$F(x) = \mu F^m(x) + (1 - \mu)F^f(x)$$
$$G(x) = \mu G^m(y) + (1 - \mu)G^f(y)$$

The copula of $H(x, y)$ is now given by:

$$C(F(x), G(y)) = \mu C^m(F^m(x), G^m(y)) + (1 - \mu)C^f(F^f(x), G^f(y)), \quad (2)$$

[11]By well-defined I imply no value judgment as to gender politics. Biological gender is here seen as a binary variable observable in data.

[12]There is evidence that gender might not be completely uncorrelated with income through differing mortality rates between male and female fetuses and the mother's circumstances and lifestyle. See Orzack et al. (2015) for an analysis of the prenatal gender ratio. Furthermore, there is evidence that the gender of the child might alter parent behavior. Lundberg et al. (2007) present evidence that father involvement and fragility of families may be affected by the gender of the child. Nonetheless, the assumption of orthogonality between gender and parent income is maintained throughout the rest of the analysis.

where C^m and C^f are the "subcopulas" for m and f.[13] Notice that the marginal distributions of the subgroups enter into the copula. The implication of thinking of the joint distribution as a mixture distribution is that changes in marginal distributions of subgroups influence the measurement of mobility. This holds regardless of whether measured are based on the joint distribution or the copula. If full gender equality is achieved, that is $F^m = F^f$ and $G^m = G^f$, then (2) collapses to the usual formulation.

As the irrelevance of marginal distributions has been touted as a great advantage of rank-based measures, this highlight an important drawback; while aggregate inequality does not influence the measure, the inequality between groups does. Equation (2) is however uninformative of how the measures are affected by this type of inequality. This is the focus of the following section.

3 Mobility measures with gender inequality

All mobility measures presented above make use of the covariance either between income in base, logs or in ranks. Using the incomes in base (or logs), the aggregate covariance is easily decomposed into male and female covariances, and the aggregate measure is thus a simple mean of the two gender-specific measures. This is not the case with rank correlations. The rank of a child, by definition, depends on a comparison group. In this setup, one may either compare the child to other children of the *same gender* or to *all* children. I will refer to the former as the within-gender rank and the latter as the total or aggregate rank.

In order to understand the importance of gender inequality for the measurement of mobility, we need a way to describe how the dependence structure between parents and children, *given* the gender of the child, affects the dependence of parent and child when using total ranks, thereby disregarding gender. In order to elicit this link, I set up a very simple data-generating process.

[13]By the following derivation: $C(F(x), G(y)) = H(x, y) = \mu H^m(x, y) + \mu H^f(x, y) = \mu C^m(F^m(x), G^m(y)) + (1 - \mu)C^f(F^f(x), G^f(y))$

I assume a data-generating process where families draw ranks rather than incomes. To simplify, I assume that the distribution of ranks is independent of gender. Gender is a random variable, s_i, and orthogonal to the parent rank. In other words, a family with a given rank will draw a female child with a rank in the female distribution which corresponds to the counterfactual rank of a male child in the male distribution. Families are therefore defined by the tuple (u_i, v_i, s_i).

Ranks are the related to income through the gender and marginal distributions:[14]

$$y = sG^{m^{-1}}(v) + (1-s)G^{f^{-1}}(v) \tag{3}$$

In this simplified setup, gender is only important insofar as the gender-specific income distributions differ, that is $G^f \neq G^m$. The aggregate income distribution of the child generation is now given by

$$G(y) = \mu G^m(y) + (1-\mu)G^f(y) \tag{4}$$

Let v^A be the *total* rank of the child, i.e. the income rank of the child when she is compared to every other child regardless of gender. Using (3) and (4), the *aggregate* rank can be expressed as a function of the *within gender* rank and the gender indicator:

$$\begin{aligned} v^A = \Lambda(v,s) = & \mu \left[G^m\left(sG^{m-1}(v) + (1-s)G^{f^{-1}}(v)\right)\right] \\ & + (1-\mu)\left[G^f\left(sG^{m-1}(v) + (1-s)G^{f^{-1}}(v)\right)\right] \end{aligned}$$

Full gender equality implies that the aggregate rank is equal to the within-gender rank, $G^m = G^f \rightarrow v = v^A$. However, this will not be the case whenever men and women have different income distributions. In most societies a rank will often translate into a lower income for women than for men.

The assumption of orthogonality between child gender and parent

[14] I need to assume invertible income distributions. This assumption may be empirically problematic as a mass point at zero is common. I return to this point in section 4.

income implies that $F^m = F^f$. A given parent rank thus corresponds to the same income regardless of the gender of the child. Under the assumption of strictly increasing cumulative distribution functions (*cdf*), there is a direct mapping from parent income to parent rank $x = F^{-1}(u)$.

Since both v_i and v_i^A are uniformly distributed, the aggregate rank correlation can be described by the covariance $Cov(U, V^A) = Cov(U, \Lambda(V, S))$.[15] While a fully analytical solution would require assuming specific distributions I show in the appendix that the first order Taylor-approximation of this covariance around a given rank, $\hat{v}$ can be expressed as:

$$Cov(U, V^a) \approx A(\hat{v}) \times Cov(U, V), \tag{5}$$

$$A(v) = \left\{ \frac{1}{2} + \frac{1}{4} \left[\lambda^m(v) + \lambda^f(v) \right] \right\} \tag{6}$$

$$\lambda^m(v) = \frac{g^f(G^{m^{-1}}(v))}{g^m(G^{m^{-1}}(v))}, \tag{7}$$

$$\lambda^f(v) = \frac{g^m(G^{f^{-1}}(v))}{g^f(G^{f^{-1}}(v))}, \tag{8}$$

where I assume equal arrival probability of sons and daughters.

The adjustment term in (5), $A(v)$, is a function of λ^m and λ^f. These two functions are likelihood ratios evaluated at different incomes. For a given rank, v $\lambda^m(v)$ is a likelihood ratio evaluated at the income corresponding to that rank in the male income distribution. Conversely, $\lambda^f(v)$ is the reciprocal likelihood ratio evaluated at the income corresponding to the rank in the female distribution. To evaluate the magnitude and sign of the adjustment term in (5) is equivalent to evaluating these two likelihood ratios.

As densities are always positive, both λ^m and λ^f are positive for all possible ranks of evaluation. This implies that the adjustment term is never negative. In other words, the total correlation will always have the same sign as the within-gender correlation. When the marginal distributions are the same, then $\lambda^m = \lambda^f = 1$. This implies that the adjustment

[15]A uniformly distributed variable on the unit interval has a variance of 1/12. The rank correlation is therefore simply a rescaling of the covariance.

factor equals one. In other words, when there is no gender inequality, the within-gender correlation equals the total correlation. What remains to be determined is the magnitude, and, maybe most pressing, when the ratio is larger or smaller than one.

3.1 Evaluation of adjustment factor

In order to investigate the size of the adjustment term, I need to make assumptions concerning the income distributions. I assume that the likelihood ratio, $\frac{g^m(y)}{g^f(y)}$, increases monotonically. Intuitively this means that the higher the income, the larger is the ratio of men to women. From this follows that the male distribution stochastically dominates the female distribution, $G^m(y) < G^f(y)$ and that densities cross only once.

An example of such a situation can be seen in Figure 1. Take two distributions, one for women (in red) and one for men (in blue). The top plot in Figure 1 displays the cumulative distribution functions of such two artificial distribution. The two functions are chosen such that they exhibit a monotonic likelihood ratio and therefore exhibit stochastic dominance and single crossing.[16]

The density functions are displayed in the middle of Figure 1. One can intuitively see that the blue density divided by the red rises monotonically with the income level and cross only once. Denote the point of crossing y^*, that is $g^m(y^*){=}g^f(y^*)$. The single crossing implies the following:

$$\frac{g^m(y)}{g^f(y)} < 1 \text{ if } y < y^*$$
$$\frac{g^m(y)}{g^f(y)} > 1 \text{ if } y > y^*$$

In other words, the male density is lower than the female density when the income is below the crossing point, and higher when income is above. Using this observation we can bound the interval, where $\lambda^m(v)$ and $\lambda^f(v)$

[16]In the empirical analysis in section 4 I show that the assumption of monotonic likelihood ratios is reasonable in an American context.

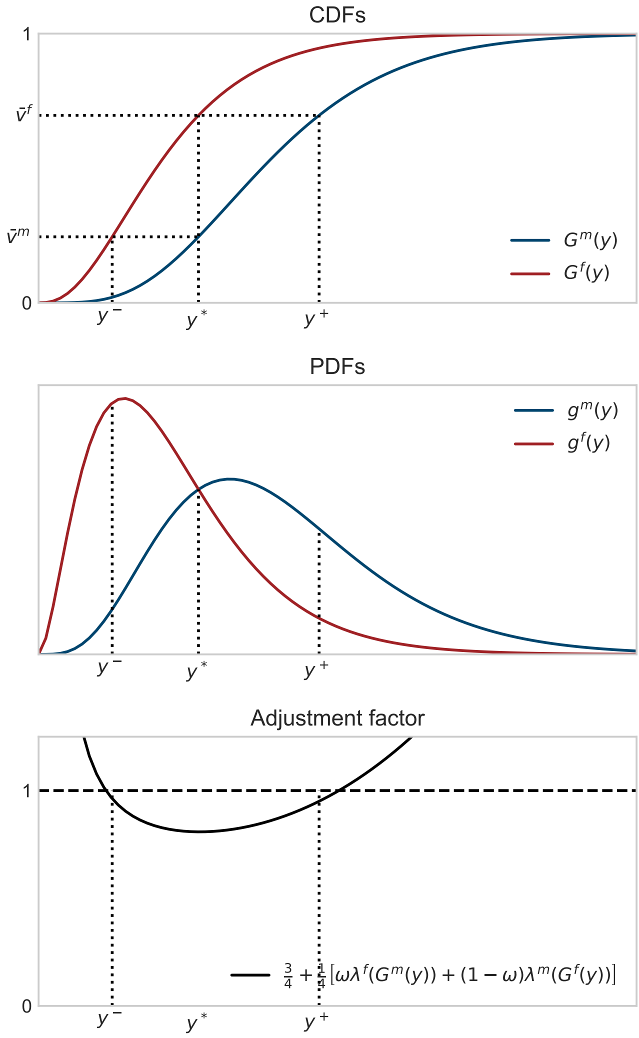


Figure 1: Range where aggregate covariance $<$ within-gender covariance

The figure depicts artificial distributions which exhibit monotonic likelihood ratios and $\frac{g^f(v)}{g^m(v)}$ is decreasing. The top figure displays the cumulative distribution functions (CDFs). From these functions one can identify $\bar{v}^m = G^m(y^*)$ and $\bar{v}^f = G^f(y^*)$ where the likelihood ratios are evaluated, where y^* is the crossing of densities. The ranks are transformed back to incomes, $y^- = {G^f}^{-1}(\bar{v}^m)$ and $y^+ = {G^m}^{-1}(\bar{v}^f)$. Evaluating the covariance at any income level in the interval (y^-, y^+) the aggregate rank correlation is smaller than the within gender correlation. With the distributions specified in this example, y^- and y^+ correspond to the 13th and the 82nd percentile in the aggregate income distribution.

are *both* smaller than one:

$$\lambda^m(v) < 1 \Leftrightarrow \frac{g^f\left(G^{m-1}(v)\right)}{g^m\left(G^{m-1}(v)\right)} < 1 \Leftrightarrow G^{m-1}(v) < y^* \Leftrightarrow v < G^m(y^*)$$
$$\lambda^f(v) < 1 \Leftrightarrow \frac{g^m\left(G^{f^{-1}}(v)\right)}{g^f\left(G^{f^{-1}}(v)\right)} < 1 \Leftrightarrow G^{f^{-1}}(v) > y^* \Leftrightarrow v > G^m(y^*),$$

where the second inequality follows from the single crossing property.

Therefore, the correction term is less than one for $\hat{v} \in \left[G^m(y^*), G^f(y^*)\right]$. Theses intervals can be read off Figure 1 in the following way. Find the crossing point, y^*, in the middle figure. Now evaluate this value in the CDFs in the top figure to find the corresponding bounding ranks, $\bar{v}^m = G^m(y^*)$ and $\bar{v}^f = G^f(y^*)$. These bounds can, in turn, be transferred back into income levels, denoted by y^- and y^+. The adjustment term is less than one for every income level within these bounds.

A calculation of the adjustment term is performed in the bottom plot of Figure 1. Note that these bounds are a sufficient condition. The adjustment term can be smaller in cases where either λ^m or λ^f are larger than one, as long as the other ratio is sufficiently small. We see, however, that the size of the adjustment term depends on income level at which one evaluates the approximation. Thus for practical application the adjustment term should be a function of an income level rather than a rank. I turn to the choice of evaluation in the next section.

3.1.1 Evaluation at an income level

The approximated relationship in Equation (5) is evaluated at a *rank* level. However, if we want to approximate the adjustment factor at an *income* level this may correspond to two different ranks, one for women and one for men. This gives rise to two different adjustment factors. In order to collapse the factors into a simple evaluation, I suggest taking the density-weighted mean, where the densities are evaluated at the income level of the overall approximation.

Denote $\bar{y}$ as the income of evaluation and $\omega = \frac{g^m(\bar{y})}{g^m(\bar{y})+g^f(\bar{y})}$. Now define the two ranks corresponding to $\bar{y}$ for men as v^m and for women v^f. We can now calculate the weighted mean of the adjustment term:

$$
\begin{aligned}
&\omega A(\bar{v}^m) + (1-\omega)A(v^f) \\
=&\frac{1}{2} + \frac{1}{4}\left[\omega\left\{\lambda^m(v^m) + \lambda^f(v^m)\right\} + (1-\omega)\left\{\lambda^m(v^f) + \lambda^f(v^f)\right\}\right] \\
=&\frac{1}{2} + \frac{1}{4}\left[\omega\left\{\frac{g^f(\bar{y})}{g^m(\bar{y})} + \lambda^f(v^m)\right\} + (1-\omega)\left\{\lambda^m(v^f) + \frac{g^m(\bar{y})}{g^f(\bar{y})}\right\}\right] \\
=&\frac{3}{4} + \frac{1}{4}\left(\omega\lambda^f(v^m) + (1-\omega)\lambda^m(v^f)\right), \qquad (9)
\end{aligned}
$$

where I have assumed equal overall shares of men and women.[17]

The density-weighted adjustment term is now a function of a single income level. An example of such an evaluation at the mean of the income distributions is displayed in Appendix Figure A.1. Taking the density weighted mean is a convenient choice as it simplifies the terms considerably. Other approaches are also possible. As long as the income level of evaluation is such that the corresponding male rank translates into an income for females below the crossing point for the densities, then $\lambda^f(v^m) < 1$. Likewise, if the corresponding female rank corresponds to an income for men above the crossing point, then $\lambda^m(v^f) < 1$. This implies that other weighted means where these conditions are fulfilled also will evaluate to a mean adjustment factor less than one.

3.2 Interpretation

The exercise above is somewhat technical and does not offer much intuition. However, it is useful to think of the gender assignment as a lottery. If the child is male, then the child rank translates into income via the male distribution and likewise for females. In a society with full gender equality, this lottery doesn't matter. In that case, income maps to the same

[17]The more general case is given by:

$$
\begin{aligned}
&\omega A(\bar{v}^m) + (1-\omega)A(v^f) \\
=&\mu^2 + (1-\mu)^2 + \mu(1-\mu)\left(\omega\left[\lambda^m(v^m) + \lambda^f(v^m)\right] + (1-\omega)\left[\lambda^m(v^f) + \lambda^f(v^f)\right]\right) \\
=&\mu^2 + (1-\mu)^2 + \mu(1-\mu)\left(1 + \omega\lambda^f(v^m) + (1-\omega)\lambda^m(v^f)\right).
\end{aligned}
$$

When I assume equal shares of males and females in the economy, $\mu = 1/2$ this term collapses to 9.

rank, regardless of gender. Any correlation between parent and within-gender child rank, therefore, goes straight through to the total rank.

This is not the case in societies with gender inequality. In that case, the random sorting matters. Take an extreme case where all women earn a lower income than any man. In that case, the random sorting may almost completely dominate the link from parent to child. In other words, given the same within-gender mobility the unequal society would be measured as much more mobile than the equal society. In Figure 1 is can be seen by the differences in CDFs. When the horizontal difference is large, the mapping matters greatly. This horizontal difference is equivalent to what Bayer and Charles (2018) calls the earnings gap. The lower the inequality the smaller the earnings gap, and therefore the less important is the gender.

4 Empirical Analysis

To illustrate the empirical relevance of my findings I calibrate the developed adjustment factors to empirical income distributions. When estimating empirical distributions, there are a number of issues such as top-coding, weighting and sampling error. As the calculations are made for an illustrative purpose, I ignore most of these things. That means that I do not regard my quantitative results as definitive, though highly suggestive of magnitudes.[18] This section proceeds as follows: First, I describe the data and estimation techniques. I then turn to the calibration of the approximated weights developed in the previous section. Lastly, I compare them briefly to results obtained from copula-simulations.

4.1 Estimation of income distributions

I estimate the income distributions on data provided by IPUMS-USA (Ruggles et al. 2017), from the largest sample available in each year from 1970

[18]Full code used to generate the results is available on my website. The program is written in Python and extensively documented. One can therefore easily change assumptions to investigate the stability of results.

to 2016. I restrict attention to individuals between the ages of 30 to 39. The variable of interest is total individual income (*totinc*). Only individuals with positive earnings are kept in the data.[19] The sample is weighted and I draw artificial samples from the original sample according to these weights. I then apply KDE to obtain densities, cumulative distributions and quantile functions for each gender in each year. I smooth the densities in order to interpolate values for incomes not present in the data.[20]

4.2 Results

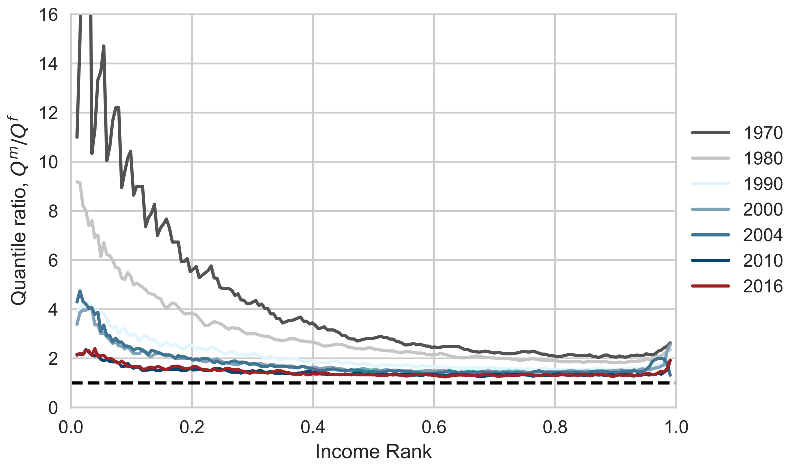
Figure 2a displays the quantile-ratios of the estimated distributions. For each rank, it shows the corresponding income for males divided by the income for women. Total lack of gender inequality would imply a flat line at one, corresponding to the black dashed line. If the ratio is above 1 the rank of a man corresponds to a higher income than the corresponding rank for a woman.[21] As can be seen, the lines are consistently above 1, indicating that the male income distribution stochastically dominates the female distribution in all years. In other words, in all years a given within gender rank will translate into a higher income for males than for females. As time has progressed, the ratios have tended downwards toward 1. This provides evidence that gender equality has improved substantially since 1970.

Figure 2b shows the estimated likelihood ratios. In order to compare the functional shape of the ratios over time the likelihood ratios are

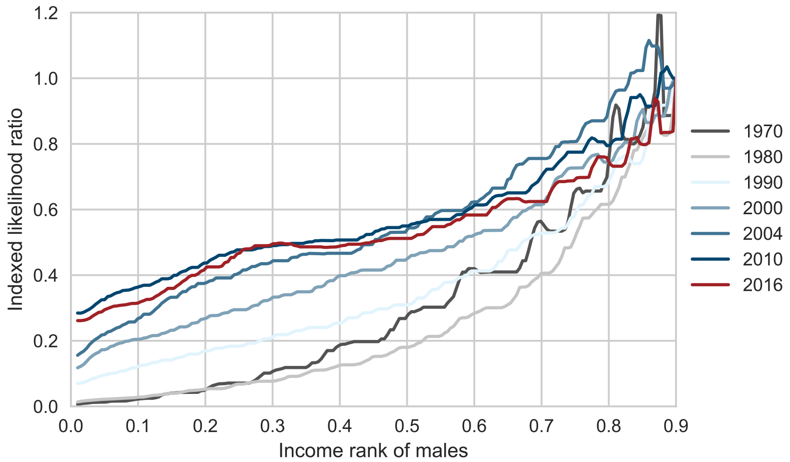
[19]This is not an innocuous restriction as it affects individuals along the external margin. The quantitative results become more dramatic is gender-inequality increases by including individuals with no income. However including zero-incomes also breaks with the assumption of smooth distributions. As the purpose of this section is to serve as an illustration, I find the drawbacks of exclusion of zero-incomes acceptable.

[20]Note, that I use income in a single year. It is recognized across the literature that this is problematic as it may be a poor proxy for lifetime income, Chen et al. (2017) and Guvenen et al. (2017) for recent discussions. However, Chetty et al. (2014b) find that the rank-correlation in the US is insensitive to the number of years over which to average income levels. The estimated distributions are only used to calculate ratios in the same year. Hence, there is no need to deflate the distributions, as the deflation cancels out.

[21]The quantile ratio is closely related to the earnings gap described in section 3.2. Formally, the earnings gap is defined as difference ${F^m}^{-1}(v) - {F^f}^{-1}(v)$, whereas the quantile ratio is given by the ratio ${F^m}^{-1}(v)/{F^f}^{-1}(v)$.



(a) Quantile ratios



(b) Likelihood ratios

Figure 2: Estimated distributions

Figure 2a shows the male quantile distribution divided by the female, and will thus take the value 1 when the quantiles are equal. I omit ratios at the very bottom, as these are very volatile due to division by values close to zero. It is evident that first-order stochastic dominance is maintained for all years. Figure 2b show the density of males divided by the density of women, i.e. likelihood ratios. For comparison, the ratios are indexed at the 90th percentile of the male income distribution in a given year and evaluated at incomes corresponding to male ranks.

evaluated at incomes corresponding to male ranks in the given year and indexed to the likelihood ratio at the 90th percentile of the male income distribution. The likelihood ratios are in general increasing for all years. This provides the empirical justification for the assumption of monotonic likelihood ratios, which was instrumental in developing the theoretical bounds on where the adjustment factor was smaller than one in section 2.[22]

4.3 Estimated adjustment factors

Using the estimated densities, I can calibrate the adjustment factor, developed in section 2. Recall, that one needs to choose an income level at which to evaluate the adjustment factor. The density-weighted mean adjustment factor evaluated at the aggregate mean and median is shown in Figure 3. The general trend confirms the intuition that the increased gender equality should in an of itself increase the total rank correlation.

From Figure 3 we see that the adjustment factor rises from 0.79 to 0.97 between 1970 and 2016. Assuming constant inter-generational mobility within gender the aggregate mobility would have *fallen* by 23 percent solely due to greater gender equality.[23] In other words, the greater gender inequality achieved in the last half of the twentieth century decreased mobility considerably. In appendix Figure B.2 I calibrate the adjustment factor for incomes corresponding to male ranks between the first and the 99th percentile. I find that it in practice matters little at which point in the income distribution the adjustment factor is calculated, as long as it is not in the extreme tails.

4.4 Simulation

The adjustment factor is developed using a Taylor approximation and one should not expect it to fit the data perfectly. Another way to approach the

[22]Again, stochastic dominance follows from the monotonic likelihood property. Thus the stochastic dominance in Figure 2a is a necessary condition for the monotonic likelihood property to be a reasonable assumption.

[23]Calculated as 0.97/0.79-1=0.23.

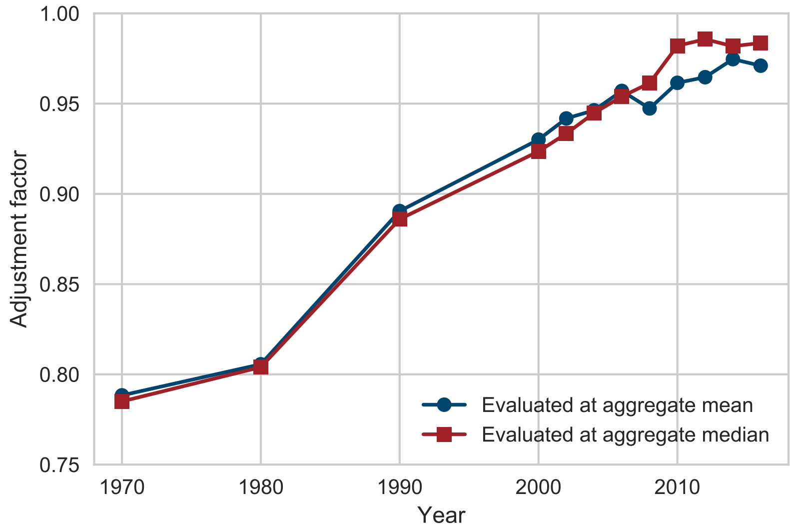


Figure 3: Calibrated adjustment factor

The figure shows the calibrated density weighted adjustment factors according to Equation (9) evaluated at the aggregate mean and median. The weights are not sensitive to choice of the income level at which to evaluate the adjustment factors. For the value of the adjustment factor for other income levels I refer to Appendix Figure B.2.

question of the significance of gender inequality is through simulation. As mentioned in section 2, a joint distribution can be described by a copula and a set of marginal distributions. The income distributions have been estimated, which leaves only the copula unknown. I assume a functional form of the copula and specifically choose a Gaussian copula, as it only contains a single parameter, which maps one-to-one to the rank correlation.[24]

The simulation exercise runs as follows: For a range of rank correlations between 0 and 1, I draw ranks for children and parents from the Gaussian copula corresponding to that rank correlation along with a random binary gender indicator. I then map the ranks into incomes accord-

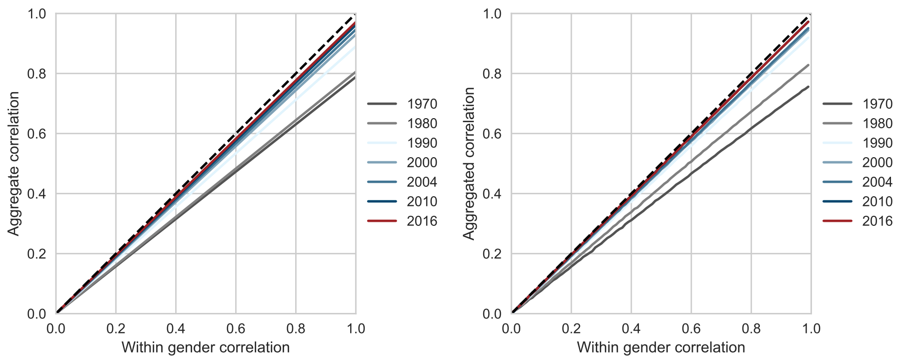
[24] The Gaussian copula has been controversial as it cannot describe tail dependence. This is relevant in mobility studies, as one often observe lower mobility in tails of the distribution. In other words, children of high very income parents are much more likely to become very high earners themselves, see Chetty et al. (2014b) and Boserup et al. (2014) for examples of such a pattern in the US, Canada and Denmark. However, as I do not concern myself with the copula itself, but rather the link between within gender and total mobility, I find these issues to be of minor importance in the present setting.

ing to the gender dummy and the estimated gendered quantile functions. The total rank correlation is then calculated as the correlation between the ranked child income and parent rank. I repeat the exercise a hundred times and take the mean of the resulting hundred rank correlation. The simulation protocol is described in Appendix C, and code is available online.

Figure 4 compares the results from calibrating the adjustment factor and the copula simulations. The linear relationship in Figure 4a is mechanical, as I just plot the within gender rank on the x-axis and the same correlation multiplied by the calibrated adjustment factor. However, I have not imposed the linear relationship in Figure 4b, where I plot the mean total rank correlations from the simulations as a function of the imposed within-gender rank correlation. The figure shows that the linear relationship between the within-gender and total rank correlation is an extremely good approximation. Appendix B shows that the approximation performs well when compared directly to the simulations using Gaussian copulas. Though the relationship is not exactly one, they follow each other closely.

Figure 4 also presents another way to interpret the importance of gender inequality for the measurement of mobility. A given rank correlation can map into many different societies. An aggregate rank correlation of 0.4 can thus map into a within gender rank correlation between 0.4 today and 0.5 in 1970. This can be seen by following a horizontal line from 0.4 on the y-axis in 4a. Furthermore, due to the proportional relationship, the larger the underlying rank correlation, the larger is the absolute difference between the aggregate and within-gender correlations.

These results show that the rank correlation is not as robust as often assumed. Interpretation and comparison of rank correlations over time are not straightforward when changes in underlying societal factors occur at the same time. Time is however only one dimension, another is space. One should expect the same issues when comparing across countries or other geographic entities where gender equality may differ.



(a) Calibrated adjustment factor

(b) Simulated relationship

Figure 4: Calibrated and simulation association between within-gender and aggregate rank correlations.

Figure 4a shows the relationship between within-gender and aggregate rank correlation with calibrated density weighted adjustment factor as described in Equation (9) evaluated at the aggregate mean. The aggregate gender correlation is simply the adjustment factor are multiplied by the within gender rank correlation. The dashed black line represent full gender equality. Figure 4b shows the simulated aggregate rank correlation as a function of a known within-gender rank correlation. The simulation assumes a Gaussian copula and follows the protocol described in appendix C. Note that no linear relationship has been imposed. The relationship with the chosen copula, therefore, seems to be almost perfectly linear. For a direct comparison between the calibration and simulation see Appendix Figure B.1.

5 Discussion and further issues

Lastly, I briefly touch upon some other issues of mobility measurement which relate to the findings of this paper. These are all important issues but are outside the scope of the present analysis.

Unit of analysis This analysis has exclusively focused on individuals as the unit of analysis. Many studies have instead used household measures. The inequality between genders is obfuscated by the coupling of male and female income, but issues with intragroup inequality remain. The weighting of single income households against dual earners will be directly affected by the dynamics presented here, though not to a full degree. Trends in assortative mating and household-level optimization also affect rank-based measures in ways, which cannot be expected to remain constant over time.[25]

Definitions of parents Another issue the present paper has not touched upon is the definition of parents. Whether parents should be defined as households or the individual parent is an issue that only complicates the interpretation of rank-based measures further. These considerations will always be present, but as long as gender is random, the definition of parents should not affect the main conclusions of this paper.

Other measures beside income Though the intuition and results in this paper are developed using income as a measure of social position, the findings may carry over to other measures such as wealth and educational attainment. The theoretical results are developed under the assumption of continuous marginal distributions, but I conjecture that the overall logic does not hinge on this assumption. However, it is naturally crucial that one is able to rank the measures in order to calculate the rank correlation.

[25] Chadwick et al. (2002) discuss these issues in the context of estimating IGEs.

Other types of group inequality An important question is whether insights of this paper translates to other forms of group inequality, the most present being ethnicity and race. Gender is characterized by the fact that all types of families have sons and daughters. This feature is not present with race or ethnicity. The interpretation of gender as a random sorting mechanism does therefore not carry over to ethnicity. To see why, note that parent income will correlate with the assignment to groups. Thus, it is perfectly plausible that the total rank correlation will be *higher* than the within-race rank correlation. This could be the case if the races are so unequal that their income distributions have little overlap *in both generations*. The aggregate and within-race rank correlations will almost surely differ but in ways different from the case of gender. Thus, while this paper documents the inequality within groups is important for the rank-correlation, the exact results are specific to gender.

Gender discrimination and equality of opportunity One can be perfectly content with using an aggregated rank measure as all the above-stated phenomena reflect economic mobility. Nevertheless, the choice of measure directly affects the interpretation and political implications. This article has shown that it is not meaningful to compare a society with large gender inequality with a society that has succeeded in alleviating gender inequality, unless one is willing to argue that, indeed, gender inequality should be interpreted as a means to increase intergenerational mobility. How to define intergenerational mobility is closely intertwined with the discussion of equality of opportunity. Gender is exogenously given and does not depend on the effort of the child. Ensuring gender equality is therefore often seen as a pathway to ensuring equality of opportunity.[26] In light of this, I claim that it makes little sense to have gender discrimination be a policy tool to ensure mobility.

[26]The question of equality of opportunity has given rise to a very large literature. For a recent general discussion see Roemer and Trannoy (2015).

6 Conclusion

The potential for confusing increased gender equality with decreased mobility shows how the interpretation of rank-based mobility measures is less obvious than it might initially appear. Copulas and ranks greatly increase the scope for research on intergenerational mobility. However, too little focus has been afforded the process of transformation of income to ranks and the corresponding changes in measures of mobility. No one correct ranking procedure exist, nor should there, but this paper has illustrated the perils of comparing mobility in societies over time and across space.

These issues are not new in the mobility literature, but rank-based measured have not yet received the careful inspection that has been afforded the measures such as the inter-generational income elasticity. This paper has provided a stepping stone towards this goal.

References

Bayer, Patrick and Kerwin Kofi Charles, "Divergent Paths: A New Perspective on Earnings Differences Between Black and White Men Since 1940," *The Quarterly Journal of Economics*, 2018, *133* (3), 1459–1501.

Berman, Yonatan, "Understanding the mechanical relationship between inequality and intergenerational mobility," 2017.

Boserup, Simon Halphen, Wojciech Kopczuk, and Claus Thustrup Kreiner, "Stability and persistence of intergenerational wealth formation: Evidence from Danish wealth records of three generations," *Unpublished Working Paper. University of Copenhagen*, 2014.

Chadwick, Laura, Gary Solon, and Laura Chadwick, "Intergenerational Income Mobility among Daughters," *The American Economic Review*, mar 2002, *92* (1), 335–344.

Chen, Wen-Hao, Yuri Ostrovsky, and Patrizio Piraino, "Lifecycle variation, errors-in-variables bias and nonlinearities in intergenerational income transmission: new evidence from Canada," *Labour Economics*, 2017, *44*, 1–12.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez, "Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States - Online Appendix," *The Quarterly Journal of Economics*, 2014, (June), qju022.

_ , _ , _ , _ , and Nicholas Turner, "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility †," *American Economic Review*, may 2014, *104* (5), 141–147.

Corak, Miles, "Do poor children become poor adults? Lessons from a cross-country comparison of generational earnings mobility," in "Dynamics of inequality and poverty," Emerald Group Publishing Limited, 2006, pp. 143–188.

Guvenen, Fatih, Greg Kaplan, Jae Song, and Justin Weidner, "Lifetime incomes in the United States over six decades," Technical Report, National Bureau of Economic Research 2017.

Haider, Steven and Gary Solon, "Life-cycle variation in the association between current and lifetime earnings," *American Economic Review*, 2006, *96* (4), 1308–1320.

Jäntti, Markus and Lena Lindahl, "On the variability of income within and across generations," *Economics Letters*, 2012, *117* (1), 165–167.

_ and Stephen P. Jenkins, "Income Mobility," in "Handbook of Income Distribution," 1 ed., Elsevier B.V., 2015, pp. 807–935.

Landersø, Rasmus and James J Heckman, "The Scandinavian fantasy: The sources of intergenerational mobility in Denmark and the US," *The Scandinavian journal of economics*, 2017, *119* (1), 178–230.

Lundberg, Shelly., Sara. McLanahan, and Elaina. Rose, "Child Gender and Father Involvement in Fragile Families," *Demography*, 2007, *44* (1), 79–92.

Meyer, Christian, "The Bivariate Normal Copula," *Communications in Statistics - Theory and Methods*, 2013, *42* (13), 2402–2422.

Mitnik, Pablo A, Victoria Bryant, Michael Weber, and David B Grusky, "New estimates of intergenerational mobility using administrative data," *Statistics of Income*, 2015.

Nelsen, Robert B., *An Introduction to Copulas* Springer Series in Statistics, New York, NY: Springer New York, 2006.

Orzack, Steven Hecht, J. William Stubblefield, Viatcheslav R. Akmaev, Pere Colls, Santiago Munné, Thomas Scholl, David Steinsaltz, and James E. Zuckerman, "The human sex ratio from conception to birth," *Proceedings of the National Academy of Sciences*, 2015, *112* (16), E2102–E2111.

Roemer, John E. and Alain Trannoy, "Equality of Opportunity," in "Handbook of income distribution," 1 ed., Vol. 81, Elsevier B.V., 2015, pp. 217–300.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek, "Integrated Public Use Microdata Series: Version 7.0 [dataset]," 2017.

Sklar, M, "Fonctions de repartition an dimensions et leurs marges," *Publ. inst. statist. univ. Paris*, 1959, *8*, 229–231.

A Taylor approximation of covariance

As in the main text, I assume that a family can be described by 3 variables:

$$(U_i, V_i, S_i),$$

where U_i and V_i are parent and child rank and S_i is an indicator variable which evaluates to one if the child is male. For simplicity, I assume that the covariance between parent and child rank is not a function of gender, $E(V_i|U_i, S_i) = E(V_i|U_i)$. This is easily changed, though the resulting equations will be less neat. Furthermore, I assume that parent income does not depend on the gender of the child, $E(U_i|S_i) = E(U_i)$. The income distributions are gender-specific and taken for granted. The translation from child rank to child income is described by the following relation:

$$y(v, s) = sG^{m^{-1}}(v) + (1 - s)G^{f^{-1}}(v),$$

where G^m and G^f are the cumulative income distribution functions of male and female children respectively. I assume these to be invertible, thereby excluding mass-points, including zero income. The *aggregate* income distribution of children is then a mixture distribution given by $G(y) = \mu G^m(y) + (1 - \mu)G^f(y)$, where $\mu = E[S_i]$.

Given these definition, the aggregate rank, v^A can now be described as a function of the within-gender rank, v. Using the *probability integral transform*:

$$v^A(v, s) = G(y(v, s)) = \mu G^m(y(v, s)) + (1 - \mu)G^f(y(v, s)) \tag{10}$$

Define $v^A = \Lambda(v, s)$. All results in this section will relate to the functional properties of $\Lambda(v, s)$.

We want to approximate the covariance between parent rank and child aggregate rank, $Cov(U, V^a)$. Due to the law of total covariance:

$$\begin{aligned} Cov(U, V^a) =& Cov(U_i, \Lambda(V, S)) \\ =& \underbrace{E\left[Cov(U, \Lambda(V, S)|S)\right]}_{\text{within gender}} + \underbrace{Cov\left(E\left[U|S\right], E\left[\Lambda(V, S)|S\right]\right)}_{\text{between gender=0}}. \end{aligned} \tag{11}$$

The decomposition shows that we can handle the description of the covariance in two separate parts. The first term is the mean of the covariance between parent rank and child aggregate rank *conditional on gender*. The second term is the covariance between means of parent and child rank for families of sons and daughters. But since I have assumed that $E(U_i|S_i) = E(U_i)$, the latter term evaluates to zero. We therefore only need to focus of the within-gender covariance.

Since S is a binary variable, it follows that (11) can be written as:

$$E\left[Cov(U, \Lambda(V, S)|S)\right] = \mu \underbrace{Cov(U, \Lambda(V, 1))}_{(i)} + (1-\mu) \underbrace{Cov(U, \Lambda(V, 0))}_{(ii)}. \tag{12}$$

As will become clear, it is easier to handle gender-specific versions of $\Lambda(v, s)$. Therefore define:

$$\Lambda^m(v) = \Lambda(v, 1) = \mu v + (1-\mu) G^f\left(G^{m-1}(v)\right) \tag{13}$$

$$\Lambda^f(v) = \Lambda(v, 0) = \mu G^m\left(G^{f^{-1}}(v)\right) + (1-\mu) v \tag{14}$$

We first focus on (i) in (12):

$$Cov(U, \Lambda(V, 1)) = E\left[U\Lambda(V, 1)\right] - E\left[U\right] E\left[\Lambda(V, 1)\right] \tag{15}$$

Insert (13):

$$\begin{aligned} Cov(U, \Lambda(V,1)) =& E\left[U\left(\mu V + (1-\mu) G^f\left(G^{m-1}(V)\right)\right)\right] \\ & - E\left[U\right] E\left[\mu V + (1-\mu) G^f\left(G^{m-1}(V)\right)\right] \\ =& \mu \underbrace{Cov(U,V)}_{\text{Within-gender covariance}} + (1-\mu) \underbrace{Cov\left(U, G^f\left(G^{m-1}(V)\right)\right)}_{\text{“Adjusted covariance”}} \end{aligned} \tag{16}$$

While the within-gender covariance is known, we cannot yet describe the "adjusted covariance". We, therefore, reformulate the second term:

$$\begin{aligned} & Cov\left(U, G^f\left(G^{m-1}(V)\right)\right) = \\ & E\left\{(U - E[U])\left(G^f\left(G^{m-1}(V)\right) - E\left[G^f\left(G^{m-1}(V)\right)\right]\right)\right\} \end{aligned} \tag{17}$$

Now perform a first order Taylor expansion of $G^f\left(G^{m-1}(V)\right)$ around $\hat{v}$:

$$G^f\left(G^{m-1}(v)\right) \approx G^f\left(G^{m-1}(\hat{v})\right) + \underbrace{\left.\frac{g^f\left(G^{m-1}(v)\right)}{g^m\left(G^{m-1}(v)\right)}\right|_{\hat{v}}}_{\lambda^m(\hat{v})} (v - \hat{v}) \tag{18}$$

Reinserting equation (18) into equation (17):

$$\begin{aligned} & Cov\left(U, G^f\left(G^{m-1}(V)\right)\right) \approx \\ & \lambda^m(\hat{v}) E\left\{(U - E[U])\left([(V - \hat{v}) - E[V - \hat{v}]]\right)\right\} \end{aligned} \tag{19}$$

Reinserting equation (19) into equation (16):

$$\begin{aligned} Cov(U, \Lambda(V,1)) \approx& \mu Cov(U,V) \\ & + (1-\mu)\lambda^m(\hat{v}) E\left\{(U - E[U])\left([(V - \hat{v}) - E[V - \hat{v}]]\right)\right\} \\ =& \mu Cov(U,V) + (1-\mu)\lambda^m(\hat{v}) Cov(U, V - \hat{v}) \\ =& \left\{\mu + (1-\mu)\lambda^m(\hat{v})\right\} Cov(U,V) \end{aligned}$$

We perform the exact same operation on (ii) in equation (12) which yields:

$$Cov(U, \Lambda(V, 0)) \approx \left\{\mu\lambda^f(\hat{v}) + (1-\mu)\right\} Cov(U, V)$$

where $\lambda^f(v) = \frac{g^m\left(G^{f^{-1}}(v)\right)}{g^f\left(G^{f^{-1}}(v)\right)}$. Reinserting into equation (12) and inserting into equation (11) yields:

$$Cov(U, V^a) \approx \left\{\mu^2 + (1-\mu)^2 + \mu(1-\mu)\left[\lambda^m(\hat{v}) + \lambda^f(\hat{v})\right]\right\} Cov(U, V) \tag{20}$$

Assuming equal share of men and women, $\mu = 1/2$, the expression simplifies to:

$$Cov(U, V^a) \approx \left\{\frac{1}{2} + \frac{1}{4}\left[\lambda^m(\hat{v}) + \lambda^f(\hat{v})\right]\right\} Cov(U, V),$$

which is identical to equation (5) in the main text.

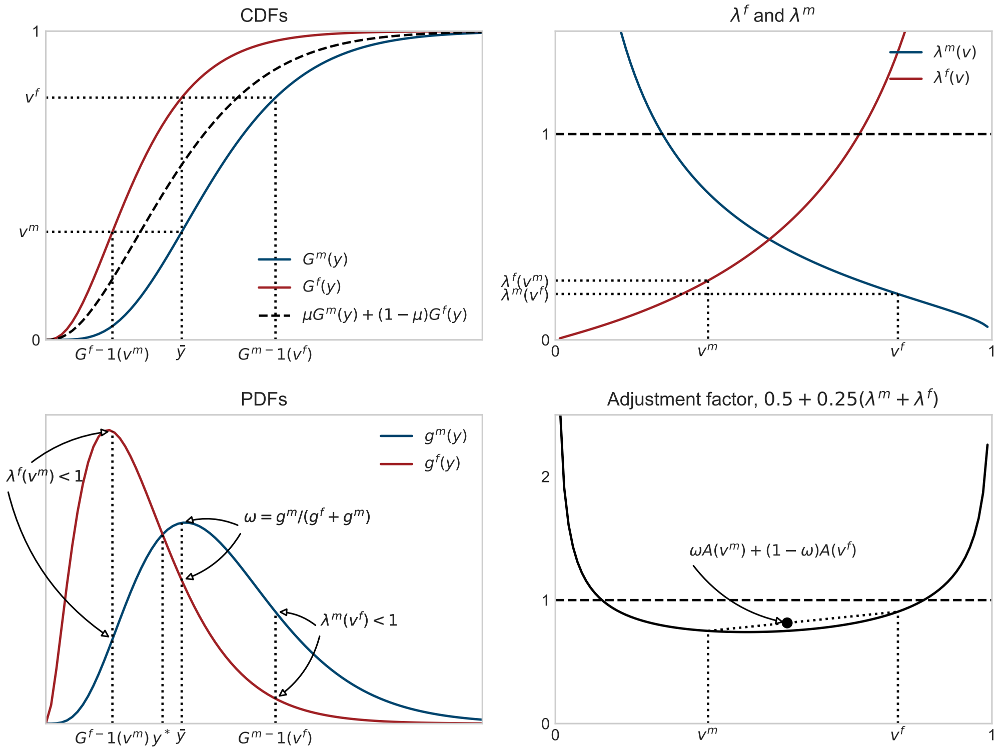


Figure A.1: Example of evaluation of the adjustment factor at the mean income

The figure depicts artificial distributions which exhibit monotonic likelihood ratios. In this example the adjustment factor is evaluated at the mean of the aggregate distribution, $\bar{y}$. The top left figure displays the cumulative distribution functions (CDFs). From these functions one can identify $v^m = G^m(\bar{y})$ and $v^f = G^f(E(y))$ where the likelihood ratios are evaluated. In this example $G^{f^{-1}}(v^m) < y^* < G^{f^{-1}}(v^f)$, where y^* is the crossing of densities. This implies that both $\lambda^f(v^m)$ and $\lambda^m(v^f)$ are below one. This can be read off the top right figure, which plots the values for λ^f and λ^m for all values of the rank, v. The adjustment factor for all values of v is shown in the bottom right figure, where the density weighted mean between the two adjustment factors are shown as a black circle.

B Additional empirical and simulation results

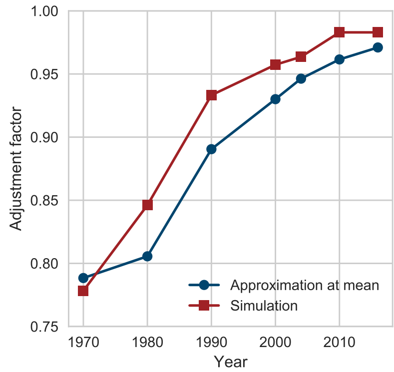


Figure B.1: Calibrated and simulated adjustment factor

The figure shows the calibrated adjustment factors according to Equation (9) in blue. It is evaluated at the aggregate mean. In red I have taken the average ratio between the total and the within-gender rank correlation across all values of the with-gender rank correaltion. As seen in Figure B.1 the realtionship between the two rank correalations is approximately linear and thus have an almost constant ratio.

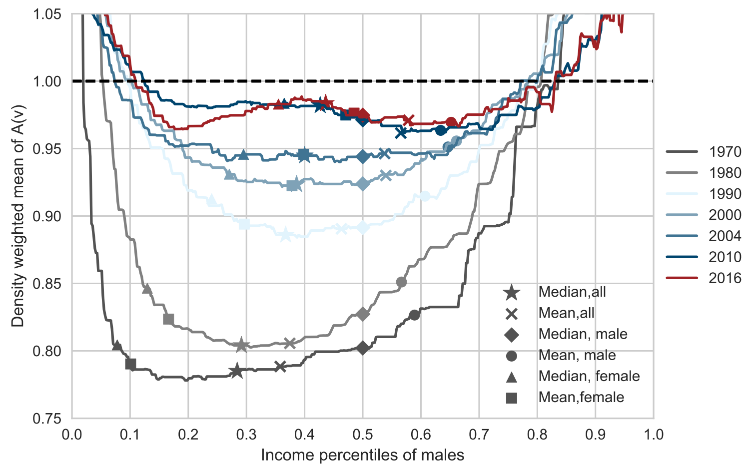


Figure B.2: Evaluation of the adjustment factor at different levels of income

The figure shows the importance of the choice of evaluation point. For each year I calculate the density weighted adjustment factor at an income corresponding to the rank on the x-axis for males. The ranks serve as a normalisation of the income distributions, such that one can compare across years. The very flat lines indicate that it matters little where the adjustment factor is approximated, less it not be in the tails. The figure also displays the adjustment factors for means and medians for the two genders and for the total income distribution. These are represented by the symbols on the line placed at the corresponding ranks in the male income distribution. By definition, the median of males are always located at 0.5 on the x-axis.

C Simulation protocol

The data is generated from a Gaussian copula, which only take one parameter which maps 1:1 with the rank correlation. Marginal distributions for parents and children have to continuous and invertible. No gender is attributed to parents and thus it is assumed that the marginal distribution of parents is uniform, $F_p(x) = x$. As described in the text children are divided into to groups m and f having different marginal distributions, G^m and G^f respectively. Let G be the income distribution containing both group m and f. The protocol for simulating is as follows:

For a given ρ repeat R times:

- Draw two vectors K_p and K_c from a bivariate normal distribution[27], $K_p, K_c \sim N\left([0,0], \begin{bmatrix} 1 & \rho_G \\ \rho_G & 1 \end{bmatrix}\right)$
- Apply the cdf of the standard normal distribution to K_p and K_c to obtain U and V: $U = \Phi(K_p), V = \Phi(K_c)$
- Depending on gender:
  - If no gender:
    - Apply quantile functions of the marginal distributions to ranks to obtain income variables:
      - For parents: $Y_p = F_p^{-1}(U)$
      - For children: $Y_c = G_c^{-1}(V)$
  - If gender:
    - Draw a binary variable, S with probability μ.
    - Apply quantile functions of the marginal distributions to ranks to obtain income variables.
      - For parents: $Y_p = F_p^{-1}(U)$
      - For children: $Y_c = S \times G^{m^{-1}}(V) + (1 - S) \times G^{f^{-1}}(V)$

Finally take the mean of the desired statistic over the R realizations.

[27] In the Gaussian copula there is a 1:1-mapping between the correlation in the Gaussian copula, ρ, and the rank correlation coefficient,ρ_s:$\rho_G = 2\sin(\rho\frac{\pi}{6})$. See Meyer (2013) for details.

Chapter 2

Defying attendance boundary policies and the limits to combating school segregation

Defying Attendance Boundary Policies and the Limits to Combating School Segregation*

Andreas Bjerre-Nielsen† & Mikkel Høst Gandil‡

Abstract

A common policy to affect student composition is to redraw school attendance boundaries. Yet redrawing only works if households comply and enroll in the designated school. Employing a novel dataset with unprecedented detail, we exploit changes over time in schools' geographic attendance boundaries to provide causal estimates of how school characteristics affect compliance with the assigned school. Households defy reassignments to schools with children from less resourceful families by enrolling in other public schools. The response to changes in school composition has a strong social gradient: resourceful households respond more to changes in school composition. We apply a boundary discontinuity design to characterize non-compliance through private school enrollment and residential relocation in the long term and once again document a strong social gradient. Our findings imply that attendance boundary policies have limited scope for desegregating schools.

*We thank Niels Johannesen, as well as seminar participants at DGPE, Danish Economic Society Meeting, Copenhagen Education Network and participants at the 8th European Meeting of the Urban Economics Association for helpful comments. We gratefully acknowledge the financial support from the Danish Economic Council of the Labour Movement(Arbejderbevægelsens Erhvervsråd) and the Danish Innovation Fund as well as the Center for Social Data Science at University of Copenhagen.

†University of Copenhagen, andreas.bjerre-nielsen@econ.ku.dk

‡University of Copenhagen and the Economic Council of the Labour Movement, mga@econ.ku.dk

1 Introduction

Policies aimed at reducing segregation in residential areas or educational institutions are common in many countries. Equalization of schools can, for instance, be achieved by physically moving people around, as with the Moving To Opportunity program, or changing the assignment of pupils to schools.[1] Assignment by school attendance boundaries (SABs), which geographically delineate school enrollment, is the most common system to assign pupils to primary schools; also known as catchment zones in the US and school districts elsewhere.[2] Local authorities can redraw geographic attendance boundaries of schools and thereby manipulate the composition of households who are eligible to enroll in certain schools. However, households may have other options than the designated school and therefore can choose not to comply. We document that Danish households exploit at least three different options to avoid enrollment into certain schools: they relocate, choose a private school, and make use of opportunities to enroll in other public schools. Along all these three margins of opting out, we find a social gradient. Well-off families react strongly to differences in school composition while marginalized families do not. The responses are large and affect the resulting peer composition of public primary schools. Redistricting will therefore not lead to the intended equalizing of student compositions across schools.

To identify household responses to school characteristics, we exploit a new dataset with geographic information on school boundaries and household residential locations for the universe of Danish children during the years 2008-2015. Attendance boundaries change over time, providing plausibly

[1]In the US laws against segregation of schools has been mandated since the Brown vs. Board case of 1954 in the US Supreme Court, see Baum-Snow and Lutz (2011). Non-discriminatory laws for housing in the US were enacted in the Civil Rights Act of 1968, see Yinger (1986). Chetty et al. (2016) investigate long-term effects of neighborhoods and reinvestigates the Moving to Opportunity program.

[2]See Monarrez (2017) on US school systems. Almost all Danish municipalities allocate students by districts. We use the term SAB and district interchangeably throughout.

exogenous variation in school assignment. We use these changes to compare households who were originally designated the same school prior to a change in boundaries but different schools after. During our observed period, a total of 191 schools, 17 percent of all schools, have parts of their district reassigned to another school. We analyze changes to schools' socioeconomic status (SES). When the new designated school has lower average SES, the enrollment in the new school drops: a one standard deviation drop in average SES implies that the compliance rate falls by 20 percent. The behavioral responses to changes in SES are not symmetric: on the one hand, a large, significant drop follows a fall in school SES in compliance, on the other hand, there is only weak evidence of higher compliance when being assigned to a school with higher SES.

We provide evidence of a strong social gradient: the response of the highest quartile is 2.5 times that of the lowest. These differences imply that the distribution of students who end up in the school depends on the initial school SES. A fall in school SES of respectively 1, 2 and 3 standard deviations compared to the originally intended school implies a drop in average SES of pupils arriving in the new school of respectively 4, 11 and 25 percent, where we assume the SES for shifted households is drawn from the population SES distribution.

We find that changes to the ethnic composition are also of importance for enrollment. The responses closely mirror those estimated using SES. Due to a high correlation between ethnic shares and average SES at the school level, we argue that, in a Danish context, we cannot meaningfully separate ethnicity and socioeconomic factors. Furthermore, we conjecture that households themselves might struggle to disentangle the socioeconomic composition from ethnicity. We are therefore unable to disentangle preferences from statistical discrimination. We finally analyze the importance of a public school-value added measure and find only weak effects on enrollment decisions.

The most common option that households use to avoid the new designated school is access to other public schools - an option provided free-of-charge through an opaque, decentralized process, which may be unfair to some households. When we identify the effects of changes to school districts, we only measure responses in the short term. Options for enrolling in a private school or relocating may, however, be limited in the short term.[3] Therefore, we turn to an auxiliary strategy to investigate long-term responses: *boundary discontinuity design* (BDD). This approach analyzes discontinuities in enrollment and relocation for households living near the administrative border between schools. By comparing school borders, BDD captures long-term responses beyond the immediate reaction to changes in school boundaries. Using BDD, we show increased differences in non-compliance through both enrollment in private and other public schools when differences in SES increases. Private schools account for approximately one third of non-compliance. Again, we document a strong social gradient. The marginal propensity to choose a private school is approximately five times larger for high-SES households compared to low-SES households. We also demonstrate that high-SES households are much more likely to move out of districts with low socioeconomic status before school age of children.

Our findings generalize to other contexts beyond Denmark for two important reasons. First, Denmark is a relatively homogeneous country with low inequality and high social cohesion. Therefore, less cohesive societies should find larger responses to changes in attendance boundaries. Second, in our sample there is *not* a positive association between school SES and school funding.[4] In a context where school finance is positively associated with school-SES, such as the U.S., there would be added incentive to sort and our estimated behavioral response would, therefore, reflect a lower bound.

[3]Using changes, we find that households do not respond along the private school margin in the short term. We speculate that the low substitution to private schools may be due to a low supply of private schools combined with the fact that they often work through waiting lists, requiring households to apply years in advance; these two facts would narrow the feasible set of private schools. Likewise, relocating quickly can be expensive and infeasible (e.g. due to commuting and financial constraints) for households.

[4]We compare public schools within the same administrative unit, i.e. a Danish municipality, where one funding scheme applies to all schools. If anything, schools with lower SES are compensated to have more resources.

Our results imply that policies that redraw school boundaries lead to systematic defiance. Forcing students into poorer districts means that a higher number of resourceful students never arrive in the designated school or possibly abandon the public school system entirely. We note that policies aimed at redistributing skills and/or opportunities by altering the structure of social interactions operate under the assumption that more resourceful peers increase one's own chances and performance (Sacerdote, 2011). Therefore, these policies, which Durlauf (1996b) refer to as 'associational redistribution', are likely to face similar behavioral responses to those that we document. Consequently, associational redistribution policies must consider the willingness to participate and the outside options of those affected.

The investigation of school assignment and compliance dates back to Coleman et al. (1966), who defined the relocation of white people from urban to suburban areas as "white flight". Subsequent work has sought to measure out-group avoidance in school enrollment (Rossell, 1975; Saporito and Sohoni, 2007; Rangvid, 2009; Bifulco et al., 2009; Riedel et al., 2010; Baum-Snow and Lutz, 2011). Our findings are consistent with their findings in that households tend to avoid schools where the ethnic composition of students differ from themselves. Papers in this literature, however, generally lack clear identification strategies to handle residential sorting. One exception is Baum-Snow and Lutz (2011), who identify responses in public school enrollment to desegregation using variation in timing of court orders.

There are several alternatives to SABs for allocating children to schools. One approach is matching mechanisms studied in a large and expanding literature, following the seminal work of Abdulkadiroğlu and Sönmez (2003).[5] Researchers have used applicant priorities over schools from truth-revealing assignment mechanism to estimate preferences for schools (Hastings et al., 2009; Burgess et al., 2015; Borghans et al., 2015; Abdulkadiroğlu et al.,

[5]The innovation in matching mechanisms is that they remove gains to strategically manipulate the assignment by submitting false preferences. These mechanisms, often known as strategy-proof, provide incentives for submission of ranking over schools that does not violate ones' true preferences.

2017).[6] The general findings relevant for our analysis are that households prefer schools that are closer to home, schools with better test performance and schools with higher average socioeconomic status (or a proxy thereof). The preferences for quality and socioeconomic composition are generally found to be increasing in households' own socioeconomic background. Our results closely mirror these findings, even though we investigate a completely different institutional setting. We complement the literature by considering new options for enrollment, as we consider private school enrollment and location decisions and show these to be relevant. By doing so, we demonstrate the importance of assignment loopholes and institutions outside the public school system, notably private schools and residential relocation, for student sorting; factors often overlooked in the school choice literature.

Our spatial identification approach is inspired by the hedonic pricing literature. A large literature employs a BDD approach to identify the effect of school characteristics on house prices Black 1999; Bayer et al. 2007; Fack and Grenet 2010; Black and Devereux 2011; Gibbons et al. 2013; Imberman and Lovenheim 2016. These studies find evidence that neighborhood composition as well as school composition and test scores affect prices. Imberman and Lovenheim (2016) show how the publicity of school performance information impacts house prices; they find no effects from school value added when controlling for peer characteristics. Our paper contributes to the understanding of BDD by showing that using difference-in-difference yields similar estimates.

The paper proceeds as follows. Section 2 gives an overview of the institutional context of primary schools in Denmark. We present our identification strategies in Section 3. We describe our data in Section 4. We perform our main empirical analysis exploiting changes to school districts over time in Section 5; this is followed by our auxiliary approach using border comparisons in Section 6. Finally, Section 7 concludes.

[6]A research literature on school preferences use surveys but it has largely been dismissed due to possible bias in reporting and a failure to account for different choice sets available to parents (Burgess et al., 2011).

2 Institutional background

We begin by describing the Danish primary school system and the options available to households with school-age children. Danish children usually start primary school in the summer of the year they turn six. The first year is grade 0, which has been mandatory since 2009. Primary school runs until grade 9 although it is not uncommon for schools to specialize in grade 0 through 6.

Municipalities who decide on the level of funding administer public schools. Public schools are free and parent co-payment is forbidden by law. The law governing Danish municipalities seeks to ensure Tiebout-competition on taxes and services but with extensive transfers between municipalities to combat inequality in funds attributable to differences in population composition. Local tax revenue, therefore, does not completely determine available funds for schooling.

Each municipality decides on the number of schools and the district boundaries.[7] Every residential address is associated with exactly one school, which we refer to as the *district* school. Children have a right to be admitted to the district school associated with their place of residence and once a child is enrolled she is not affected by future district changes (disregarding mergers and closures).[8] The municipal council is free to choose its priorities when constructing the districts. Changes in districts are common and some municipalities use redistricting as a way to manipulate student body characteristics, especially in large urban areas such as the Copenhagen metropolitan area, Aarhus and Odense.[9] Parents have a right to have their child admitted to other schools than the district school if the desired school has enough capacity. This is usually defined by a cap on class size and the total number

[7]Each district has one school; thus there is no distinction between catchment areas and school districts in Denmark.

[8]In recent years some municipalities have merged smaller schools into one organizational unit to lower costs.

[9]We interviewed the responsible administrators in the municipality of Copenhagen and consulted administrative texts to verify this to be the case.

of pupils in a school.[10] Children are free to change school during the school year. This creates the possibility that an initially oversubscribed school may become accessible for outside-district children if a family moves away.[11]

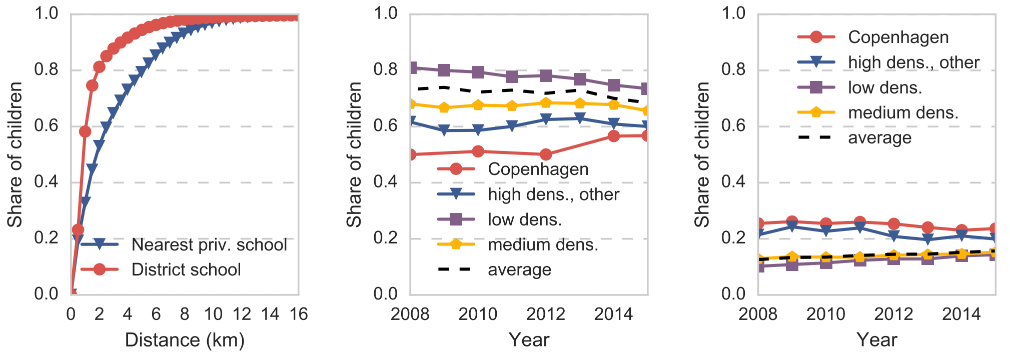
Private schools receive funds from the government covering on average 75 percent of their costs, while parents cover the rest. This is similar to voucher schools in the US education system. Danish private schools are free to set their own price, and a typical monthly fee will be around 130-270 euros a month per child with discounts for siblings.[12] These schools are free to choose who to admit and thus have no SAB. Popular private schools have waiting lists and parents can sign their children up for these very early in the life of their child. Parents, therefore, cannot be sure to exploit private schools as outside options as this is contingent on being on a waiting list and being admitted. Anecdotal evidence suggests that private schools, especially in urban areas, tend to be vastly oversubscribed.

The supply of private schools in Denmark is evenly distributed geographically and most areas have a private school nearby. The distribution of distance to nearest private school in Figure 1a shows that around 42 pct. of children aged 7 have a private school within 2 km and 85 pct. have one within 10 km. The overall enrollment in private schools increased from around

[10]The municipality can delegate to the school principal the authority to suspend the right of outside-district children to be admitted to a certain class or year in a school. Generally, a school class must not exceed 28 at the beginning of the school year, although under special circumstances the municipality council can allow classes to reach a maximum of 30. The municipality can decide on a separate class size limit for which pupils from outside the school district can no longer enter. If a school receives more applications than its capacity the children outside the district should be admitted according to objective criteria. The Danish Ministry of education recommends distance and sibling preference as such criteria. See guidelines on the Danish school choice system https://www.uvm.dk/folkeskolen/fag-timetal-og-overgange/skolestart-og-boernehaveklassen/frit-skolevalg

[11]No centralized mechanism exists for the transitions between schools and the chance of admission depends on the timing of the request to move. It is, therefore, possible for parents to increase the opportunities for admission by repeatedly contacting the desired school. We cannot follow this process in our data.

[12]We have been unable to locate a central registry of prices, and our estimates are therefore based on data collected on the webpage of the private schools association; `https://privateskoler.dk/skolerne/liste-over-skolerne`.



(a) Distance to school (b) Share in district school (c) Share in private school

Figure 1: School distance and enrollment

The figures depict various statistics for distance and enrollment. Figure 1a shows the cumulative distribution of distance to district school and nearest private school. Figures 1b and 1c plots the annual share of children enrolled respectively in the district school and in a private school. The sample consist of all children at the age 7 between 2008 and 2015. For enrollment, the density measures are: low density, less than 1000 per sq. km; medium density, between 1000 and 5000 per sq. km, and; high density, more than 5000 per sq. km.

12.5 pct. in 2007 to 16.4 in 2016. Figures 1b and 1c shows a breakdown of school enrollment in the district school and in private school by population density. As is evident from these figures, urban density is an important determinant of enrollment. Enrollment in the district school is around 75 percent nationally but as 50 percent for the Copenhagen area, which is reflected in a corresponding larger private school enrollment compared to the national average.

3 Methods

In this section, we present our approach to identify households' compliance behavior as a function of their district school. We begin by defining the options available for the households to opt-out of the school assignment. We then proceed to discuss challenges to the identification of behavioral effects and our econometric approaches. We finish of with an empirical example of how our strategies are implemented in practice.

Options of households: A household with a child ready to enroll in primary school has several options when deciding on the school in which to enroll. The household may simply choose to enroll the child in the district school, thereby complying with the assignment mechanism. We describe this by a binary variable, denoted $comply$. This option is guaranteed by law and therefore always available. If the district school is deemed unattractive households may choose different ways of opting out. We group these responses into three binary variables: move to another district ($move$); enrolling in another public school ($othpub$) and enrolling in a private school ($priv$). Summing gives us the following identity:

$$comply = 1 - (move + othpub + priv), \tag{1}$$

where it is implicit that we regard movers as one, regardless of whether they end up attending the district school in their new location. Each of the elements in (1) correspond to dependent variables and we use this identity to decompose responses.

To identify household responses, we take a reduced-form approach. This implies that we do not model preferences and costs explicitly. Instead, a child's school enrollment can be seen as revealed preferences of the parents, i.e. opting into the district school or not. In order for a child to enroll in a private school a number of conditions have to be met: 1) households must prefer the private school over the associated district school; 2) the child must meet admission criteria, and; 3) the household must be able to afford enrollment. Conditions 1 and 2 also apply when choosing other public schools. As a consequence we can be sure that whenever we observe a child enrolled in a non-district school, this school must be preferred to the district school.[13] We do, however, not observe those that would want to escape the district school

[13]There may in practice be cases where a non-conforming child is urged on by authorities to move to another public school. We do not expect these cases to influence our results.

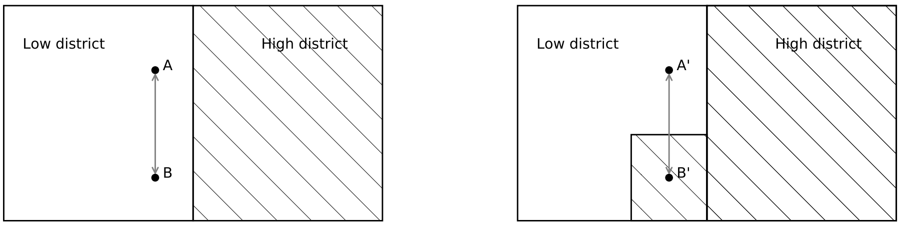
but are unable to, either due to financial constraints, over-subscription or inability to transport oneself. Whenever we refer to our estimates as reflecting preferences, it, therefore, comes with the caveat that preferences cannot properly be separated from costs and constraints.

In order to identify household responses to school characteristics, we need to address the fact that households do not choose their residential location randomly. This implies that there is sorting by households characteristics across the school boundaries. Some of these characteristics are observable and possible to control for. Yet, other characteristics are not observed and may pose a threat to the identification of behavioral responses to school characteristics. To address these concerns, we rely on two approaches which we describe below.

3.1 Main strategy: Differences-in-Differences (DiD)

Our main approach exploits quasi-exogenous redrawing of school boundaries in a Differences-in-Differences (DiD) setup. We use the fact that the redrawn school boundaries imply a different assignment of households to schools. Our method is to compare households, who used to live within the same attendance boundary but after reassignment differ in school association. We measure the enrollment difference between these two groups before and after the reassignment to schools.

We explain the method with an example. Imagine two households, A and B, at year -1. As depicted in Figure 2a the two households have chosen to locate in the same SAB. The two households may differ and therefore may make different choices with regards to school enrollment. At year 0 the municipality chooses to redraw the district borders as depicted in Figure 2b. This will not affect household A and B as they have already enrolled their child. But their neighbors A' and B' have younger children and are therefore affected by the redistricting. If the change is unexpected we may assume that, in absence of the redistricting, the behavior of household A would have been comparable to household A' and likewise for B to B' (except for a com-



(a) Comparison before change in attendance boundaries (b) Comparison after change in attendance boundaries

Figure 2: Illustration of main identification strategy

The two figures visualizes the main approach of identifying household responses to variation in school districts. Both figures illustrate two bordering school districts where the left-most district exhibit a 'low' measure of district school characteristic. From Figure 2a we see that before the redrawing of boundaries both household A and B are in the 'low' district. After the change, the area where household B lives is reassigned from the 'low' to the 'high' school, see Figure 2b. Our method is to measure the differences in the actions of households A' and B' to differences between households A and B.

mon trend). Under this assumption, we can identify the behavioral change in outcomes as $E[y_{B'} - y_{A'}] - E[y_B - y_A]$, which is a standard difference in difference estimator.

We note that by defining our measure of interest as a function of school characteristics we are implicitly assuming a time profile of mechanisms linking school characteristics to household behavior. We assume that school characteristics appear *earlier* in a causal chain than other factors such as real estate prices and allocated school resources. An example; assume household respond to a wealth effect from rising prices due to a boundary change. In order for our estimates of marginal effects to be valid, we need to assume that this wealth effect is due to the change in school characteristics. In this setting, the wealth effect is, therefore, a mechanism for the causal link between school characteristics and household school choice.

We limit our analysis to addresses being shifted between active schools and thus we do not use school closures. We do this in order to avoid a common pattern in rural areas where private schools close (and then local communities reopen the school as a private school in the same location).[14] Therefore, we do not want to attribute the household response from school closures to school characteristics.

A drawback of the DiD-approach is that the estimates only capture short-term responses after redrawing boundaries. Households may, however, be restricted in their options in the short term. If a nearby private school is oversubscribed it will not be an option for parents who want to enroll their child with short notice. Likewise, moving can be a lengthy and costly process and parents may therefore not react immediately. We therefore complement our main approach with an auxiliary strategy to investigate long-term responses to district school characteristics.

3.2 Auxiliary strategy: border discontinuity design (BDD)

We move on to describe our auxiliary approach where we compare neighbors associated with different schools. We rely on Boundary Discontinuity Design (BDD), a method first proposed by Black (1999), to evaluate how housing prices reflect school characteristics. Figure 3 provides an example of our approach. Household A and household C live very close to each other but on either side of an attendance boundary. The local school is not the only consideration when choosing a location of residence, other factors play a role such as access to labor markets, local amenities and location of relatives. Assume that these local factors do not depend on the characteristics of the associated school. We can then compare A and C to elicit when and how households choose to opt-out of the district school as a function of district school characteristics. By our assumptions, we attribute any discontinuous difference in probability of opting out to a difference in school characteristics.[15]

[14]An example of this dynamic can be seen here (in Danish): `https://www.folkeskolen.dk/18759/kommune-tjener-paa-at-foraeldre-aabner-friskole`.

[15]The boundary discontinuity design is very similar to regression discontinuity design. The difference is that in geographic space we have two running variables, a Northern and an Eastern coordinate. As the coordinates have the same scale they are easily projected

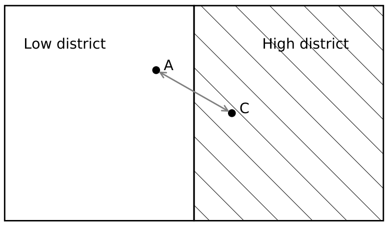


Figure 3: Illustration of auxiliary approach: Boundary Discontinuity Design

The figure displays our auxiliary approach which compares households' opt-out responses in two bordering school districts. In this figure, the left-most district has a 'low' measure of school characteristic compared to the right-most. In the figure, household A lives on the 'low' side while household C lives on the 'high' side. This approach exploits the discontinuous difference in assigned school characteristics at the boundary to see how they relate to enrollment choices.

The BDD approach is static in the sense that we do not exploit intertemporal variation in boundaries and location. This is, therefore, best seen as a description of long-term (i.e. steady state) behavior. As noted we can therefore expect the options by which households can opt-out to be less restricted. Compared to our main approach using changes in boundaries over time, the BDD design puts fewer restrictions on which data points can be used for computing responses and therefore increases the statistical power which allows us to estimate heterogeneity in responses. This increased power, however, is a function of more restrictive assumptions; household A and C should indeed only differ in that they are assigned to two different district schools.

If this assumption is not valid we see two general narratives by which our estimates might be biased. Assume that households who value school quality for unobserved reasons anticipate the schools' characteristics and never locate within an attendance boundary for a low-quality school. By comparing households across boundaries, the households on the side with lower quality will exhibit lower preferences for good schools than those house-

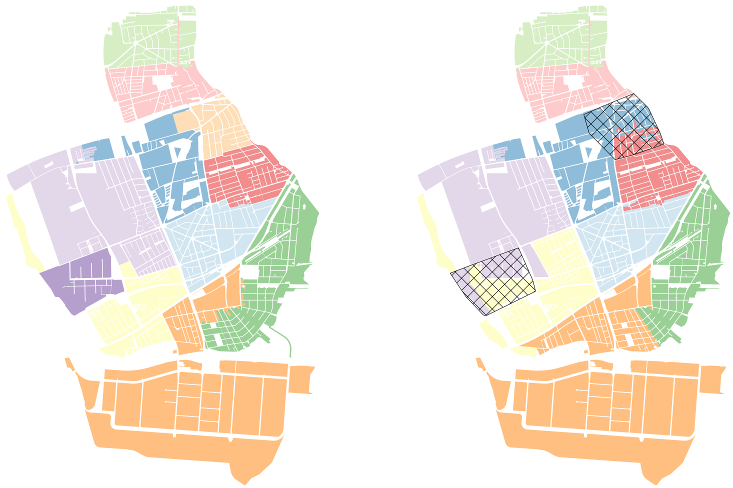
onto a line as the Euclidian distance to the border. In other words, two running variables are collapsed into one.

holds who chose to locate on the high side. The missing counter-factual households will lead to an underestimation of behavioral responses in the BDD-approach. This effect may, however, be counteracted. There is ample evidence that house prices correlate positively with school quality. See Gibbons et al. (2013) for a brief review. Bjerre-Nielsen and Gandil (2018a) provide evidence of such dynamics in a Danish context. Assume that some households have strong preferences for private school. These households can get a house for a lower price in the less attractive school district. But in the counter-factual case, where these households were living in a better district, they would still choose private school. This would have the inverse effect, that is, we would overestimate the behavioral response when comparing households across borders. We can therefore not sign the possible bias.

3.3 Empirical example

In this final sub-section, we explain with a brief example, how we identify behavioral responses in a real institutional setting. The context is Hvidovre, a municipality in the Copenhagen metropolitan area, with around fifty thousand residents. The two maps in Figure 4 illustrates how the SABs were redrawn between 2011 and 2012 where two public schools closed down.[16] To allocate the children living within the attendance boundary of the now-closed schools, other boundaries needed to be changed. This implies two kinds of variation. Firstly, households who thought their children would attend a school which by then was closed had to attend another school. These are the households with children who in 2012 lived in the hatched areas of Figure 4b. Secondly, other households were reassigned to new schools that they had not expected although their originally designated school did not close. An example of this is the change from the green to the orange for some areas in the South. We only use this latter source of variation in our

[16]The two closed schools are Sønderkærskolen, the orange district in the North-East, and Enghøjskolen, the Western dark purple district.



(a) School districts, 2011 (b) School districts, 2012

Figure 4: District changes in the municipality of Hvidovre

The figures depict the school districts in the municipality of Hvidovre in the autumn of 2011 and 2012. The hatched areas in 2012 show the convex hull of two closed schools. In order to enroll students who would live in districts of the now-closed schools, a range of other changes was made. Only the latter changes are used for identification. See section 4 for a description of the district data. Some areas differ from the official documentation. These areas are mostly not populated but some measurement error occurs. The map is constructed by merging addresses on to official geodata. In the analysis, we use addresses directly to bypass mismeasurement of geographical entities. The district polygons are only used to measure distances to borders.

Difference-in-Difference analysis. In the auxiliary approach, BDD, we exploit discontinuities in school characteristics at boundaries in both Figures 4a and 4b. Importantly we only compare within municipalities, which ensure equal levels of taxation and overall school funding.

4 Data and measurement

Our sample is based on Danish registry data for year 2008-2015.[17] From Statistics Denmark we obtain detailed information on household income, education ethnicity, educational enrollment and test scores. We link these records to a detailed geographical information on over 95 percent of households in Denmark.[18] We link this data with school district data obtained from records in the CPR-vej-register. These are reported by the municipalities themselves and are not verified by Statistics Denmark. We clean the district data and merge them unto the place of residence of households in the registers.[19] We calculate the distance to the boundary from the centroid of geographical polygons, in which the household lives.

Our data is sampled from all children who are observed at age 5 and enrolled in primary school at age 7 during our sample period (irrespective of siblings, pre-school institution choice etc.). We focus on 7-year-old children as this age captures the earliest point in time by which we expect all children to have enrolled in primary school (some parents defer enrollment until their children turn 7).

We use data on outcomes, i.e. enrollment of the child and residential location of the household, for the year the child turns 7. We require observations at age 5 in order to measure school characteristics and household covariates before the children possibly experience changes to their school district and/or enroll in primary school.

For ease of interpretation, we construct a socioeconomic index (hence-

[17]The data goes further back but the data quality on addresses, and thus geographic data, suffers from a break in 2007 when Denmark implemented a large reform of municipalities.

[18]We have constructed a set of polygons such that k-anonymity of the households is maintained, see Bjerre-Nielsen and Gandil (2018b) for details. Software is available at GitHub; `https://github.com/abjer/private_spatial_dk`

[19]For manipulation of data we have made extensive use of open-source Python libraries. Among others we have used Pandas, Scipy, Scikit-learn and NetworkX for data structuring, see McKinney (2010); Jones et al. (2001–); Pedregosa et al. (2012); Hagberg et al. (2008); GeoPandas and Shapely for GIS-data manipulation, see Gillies et al. (2007–); Matplotlib, Statsmodels for respectively plotting and regression models, see Hunter (2007); Seabold and Perktold (2010).

	N	mean	median	std	binary
Socioeconomic status [SES]	578,903	0.50	0.50	0.29	N
Employment parents, min. [EMP]	581,457	0.74	1.00	0.44	N
Income rank parents, max. [INC]	581,449	0.62	0.67		Y
Non-western [NW]	581,457	0.13	0.00		Y
High cycle educ. parents, max. [HCU]	578,903	0.19	0.00		Y
No educ. after prim. school, min. [NE]	578,903	0.07	0.00		Y
Number of parents	581,457	1.82	2.00		Y
Housing contract: rental	581,457	0.21	0.00		Y
Housing contract: coop.	581,457	0.04	0.00		Y

Table 1: Descriptive statistics for children and their households

The table presents the mean, std. deviation and count of observations for variables that we employ in the analysis as covariates for matching, for modelling or in to compute the SES index (see Appendix A.1).

forth, SES). We define this as the first component from a principal component analysis (PCA) on income rank, an employment dummy and dummies for long-cycle education. We then rank the resulting indicator such that it is uniformly distributed and bounded on the unit-interval. Our socioeconomic index increases with income, employment and high cycle education as expected. Appendix A.1 describes this SES-index in detail. Ethnicity is not part of the PCA analysis and is investigated separately. We measure ethnic background with a dummy for being a non-Western immigrant, descendant or child of descendants (up to the third generation).[20]

We focus on three school measures: ethnic composition measured as the share of non-Western immigrants (abbreviated NW); average socioeconomic status (SES), and; school value added (SVA). We measure both NW and SES in the data. We calculate averages of all students enrolled in a school for each year. For SVA, however, we make use of official measures calculated by the Danish Ministry of Education. SVA is calculated using a version of the empirical Bayes estimator. The outcome is (uncentered) grades in the final exam in grade 9, which corresponds of the final year of secondary school.

[20]In order to be a descendant both parents must be non-Danish. Same goes for children of descendants. Thus, one Danish parent is sufficient to be of Danish descent.

	# obs.	Mean	Std. err.
Non-Western share	9332	0.09	0.13
SES index (average)	9332	0.47	0.10
School Value Added	3714	0.04	0.34

Table 2: School descriptives

The table presents a descriptive statistics for schools, where each school is represented once per year.

The measure is calculated every year for new cohorts. The controls do not include pre-school test scores and therefore may suffer considerably from omitted variable bias. Furthermore, urbanization is not taken into account and we suspect the presence of substantial unobserved sorting. The measure is volatile with a year-on-year correlation in subject-institution value added of 0.3. We use SVA measured on grade averages as our measure. Importantly, these measures are available publicly on the Ministry website and are therefore plausibly part of the information set of parents. This also allows us to ignore measurement errors of the SVA in our estimations, as we take these as given from the point of view of households.

An important factor for the choice of school is the geographical distance (see Abdulkadiroğlu et al. (2017) for an example.) We calculate Euclidean distances from the place of residence (centroid of resident-polygon) when the child is five years old to the original and the new school and take the difference. When a district changes, the distance changes differently for each household depending on the place of residence. Therefore, contrary to our other school measures we have more variation in distance changes than in our other school measures, which are the same for all households in the district. Descriptive statistics for schools are presented in Table 2.

5 Main approach: Changes in attendance boundaries

We now move on to the analysis of the behavioral responses to local school characteristics.

The administrative procedures of changing attendance boundaries differ between municipalities. The changes are usually announced no more than a year before they occur, usually in the spring before the school years beginning in August.[21] Proposals for changes are usually made by administrative staff at city hall and are subject to the confirmation by the city council. The changes most often occur due to changing demographics, which induce shifts in the demand for primary schooling.[22] For all residential locations, we record changes in formal school affiliation and the year the change occurs. We restrict our attention to addresses, which experience a single change or no change in our data. For each address, which experiences a change in affiliation, we calculate the time span in years between the current year and the year of the change. We record the outcome of children the year they turn 7. We then find the address of these children at age 5 and merge it onto the attendance boundary data. A temporal difference to the SAB-change of zero implies that the change occurs between the ages of 6 and 7 of the child. A distance of 1 means that the change occurs at the ages of 5 and 6. We exclude all attendance boundaries wherein no address is shifted at any point in our data.

[21]We have interviewed responsible authorities in the municipality of Copenhagen as well as gone through public documents from other municipalities to understand the process.

[22]When a proposal of a change is made citizens and schools may voice concerns, which sometimes turns into heightened local political tension. Anecdotal evidence suggests this is mostly due to the closing of schools as opposed to tiny changes around borders. Therefore, enacted changes in this context might not be completely random as some areas are likely more difficult politically to manipulate due to a politically strong citizenry. In order to ensure exogeneity, we focus on transfers between existing schools. We exclude mergers and closings of schools by requiring that both schools involved in an exchange exist before and after.

5.1 Positive and negative shocks

We estimate household responses to changes in school affiliation in a difference-in-difference framework. To capture the effect from characteristic changes separate from "pure" effect from a surprise change we construct a treatment indicator for all addresses, which are shifted. We categorize the school changes into three groups depending on the change in district school characteristics; those experiencing a positive, negative and a "negligible" change. The latter category is taken as reference. We estimate the responses using OLS specifications of the following kind:

$$\begin{aligned} Y_{iast} &= \alpha_T T_a + \alpha_- T_a^- + \alpha_+ T_a^+ \\ &+ \sum_{k=-4, k\neq -1}^{4} \left[(\beta_T^k T_a + \beta_-^k T_a^- + \beta_+^k T_a^+) \times K_{iat} \right] + \mu_{st} + \varepsilon_{iast}, \end{aligned} \tag{2}$$

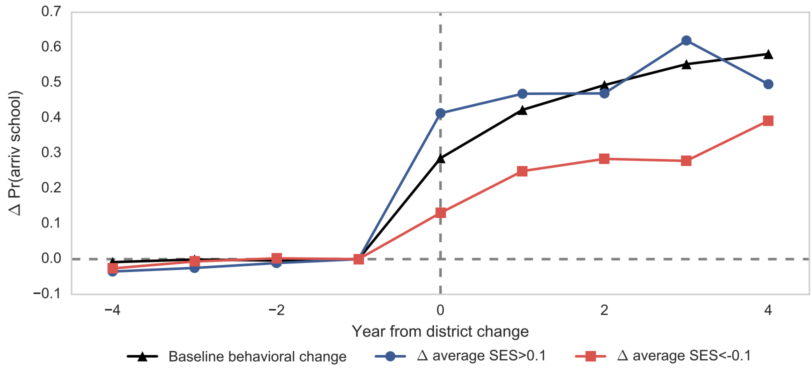
where Y_{its} is the outcome of interest for child i aged 7 at time t living at address a at the age of 5 in original SAB s. Define the dummy variables for the positive and negative treatment as well as a general treatment indicator, respectively denoted T_a^- and T_a^+ and T_a. K_{iat} is a set of dummies for the time gap between the change of association and the year the child outcome is measured. We center our estimates at the year before the association change. We are interested in comparing differences across addresses within the same year which share the old school association and implement this by including a fixed effect for the original SAB s at time t. The effects of interest are the coefficient on the interactions; β_T^k, β_-^k and β_+^k, where coefficients for $k < 0$ serve as placebo tests.

The specification in Equation (2) implies that β_-^k and β_+^k are interaction terms. In other words, the parameters describe how the mean effect of a change in SAB is affected by the change in characteristics. As mentioned we have multiple measures of school characteristics. In this section, we present our graphic results from using average SES as our school measure. We perform the same analysis for changes to share of non-Western heritage and school value-added in Appendix B.

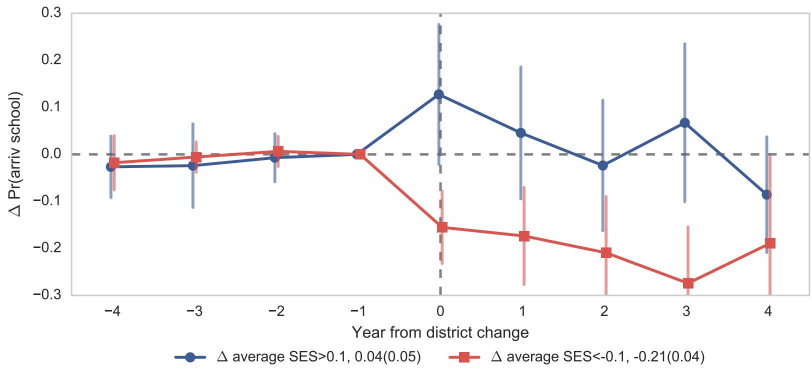
We begin by estimating our model for enrolling in the assigned school after a change in school SAB.[23] The black line in Figure (5a) depicts the change in the probability of enrolling in the new school as a function of time from the SAB change. The probability is compared to the households from the same original SAB, but who were not transferred. If redrawing attendance boundaries is an effective instrument, the probability of enrolling in a newly assigned school should rise discontinuously at the time of change. This is clearly the case. The probability of enrolling increases by almost 30 percentage points the first year and continues to rise to around 50 percentage point, which is well within the rage of common enrollment rates. The red and blue lines depict the enrollment when the change in average SES is numerically larger than 0.1, representing a standard deviation of the school level SES distribution. If households experience a positive change in school-level SES of more than a standard deviation, the compliance rate rises by around 10 percentage points in the first year, which however is insignificant as seen figure 5b. Four years after the change compliance rate converges to the baseline, which is most likely due to the sorting over time as incoming families with children know the new school association in later years. If, on the other hand, the average SES *falls* by a standard deviation the average compliance over the years is 21 percentage points lower than average enrollment rate and does not seem to converge over time. This implies that there are lasting falls in compliance for areas where the newly assigned school has a weaker socioeconomic composition. Together with flat pre-trends, this provides evidence of a causal relationship between school characteristics and compliance with the school assignment mechanism.

Larger changes cause larger responses The change of +/- 0.1 in average SES is somewhat arbitrary. We may expect that larger changes will lead to

[23] For the non-shifted addresses within a SAB, we assign arrival school as the arrival school of those that are shifted. A few SAB experience exchanges between multiple other SABs. In these cases, we assign the closest possible arrival school to the non-shifted addresses.



(a) Change in probability or enrolling in new school



(b) Excess change in probability or enrolling in new school as a function of change in SES

Figure 5: Compliance response to change in SAB by school characteristic change

Figure 5a display changes in estimated compliance rates based on the model in (2). The black lines depict the estimated β^k_Ts, while the blue and red line depict $\beta^k_T + \beta^k_-$ and $\beta^k_T + \beta^k_+$ respectively. Figure 5b displays the interaction terms, β^k_- and β^k_+, along with 95-percent confidence intervals. The parameters represent the difference in the likelihood of enrolling in the new district school, when the average SES at a school level changes, relative to the average arrival probability following a district change. The dependent variable is binary and equals one if the child is enrolled in the district school at age 7 based on the district at age 7 for the address at age 5. The y-axis denotes the excess probability of enrolling relative to baseline. The model is estimated with “origin-SAB”-year fixed effects. Standard errors are clustered on origin SAB level. Results are centered at the year before the SAB-change. Estimates from a simple before-after-DID are reported in the legends of figure 5b.

larger responses.[24] To investigate this aspect, we collapse our regression into the following:

$$
\begin{aligned}
Y_{iast} = {} & \alpha_T T_a + \alpha_- T_a^- + \alpha_+ T_a^+ \\
& + \left(\beta_T T_a + \beta_- T_a^- + \beta_+ T_a^+\right) \times Post_{at} + \mu_{st} + \varepsilon_{iast},
\end{aligned} \tag{3}
$$

where $Post_{at} = 1$ if the household is observed after the SAB-change. This is a classic two-period DiD-estimator with heterogeneity in the treatment intensity. By letting the limit for which we categorize the change to be positive or negative vary, we can elicit response size as a function of the school characteristics change. We estimate this model for addresses which are within two years of the change (-2 to 2) and those who do not experience a change. The result for overall compliance is shown in Figure 6. For a fall in average SES the absolute response is monotonically increasing. Thus, larger falls in SES entail lower compliance. When average SES rises, however, we do not observe the same functional relationship - the estimates are small, stable and insignificant. In other words, households respond to a lowering of socioeconomic status of schools but do react to to the same degree when school SES increases. This asymmetry is also found when using other NW-share as school measure.

5.2 Continuous differences in school characteristics

In the previous sub-section, we have shown that changes in compliance are a function of changes in average SES and that non-compliance occur primarily through enrolment in other public schools. To further quantify the responses, we now employ a two-period difference-in-difference model with *continuous* treatment. This model allows us to compute heterogeneous effects in socioeconomic status and control for changes in distance from home to school.

Let $\Delta Q_{ss'} = Q_{s'} - Q_s$ be the difference in a school characteristic between

[24]This would be the case if the marginal response to school quality was constant.

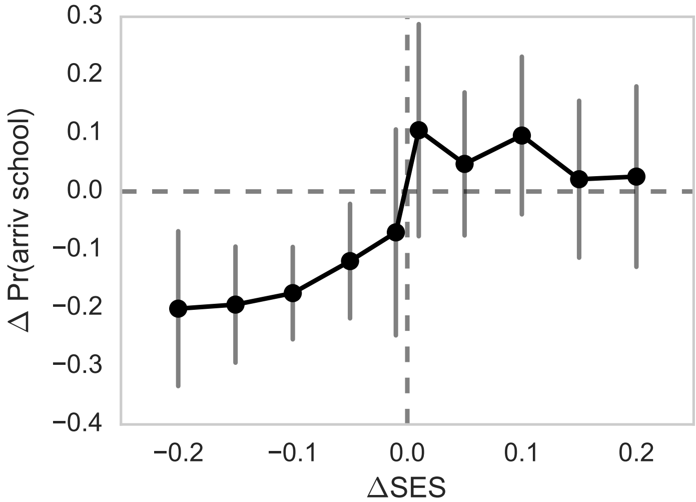


Figure 6: DID-estimate as a function of change size in average SES

The figure displays the interaction terms, β_- and β_+, along with 95-percent confidence intervals for models with different thresholds for defining a positive or negative change. The threshold is given on the x-axis and the corresponding DiD-estimate is shown on the y-axis. The positive and negative changes are estimated jointly, which implies that estimates with the same distance from 0 on the x-axis stem from the same estimation. The models are estimated with "origin-SAB"-year fixed effects. Standard errors are clustered on origin SAB level.

schools s' and s recorded the year before address a experience the change.

$$Y_{iass't} = \alpha_0 T_a + \beta_0 \Delta Q_{ss'} + \alpha_1 T_a \times Post_{at} + \beta_1 \Delta Q_{ss'} \times Post_{at} + \mu_{st} + \varepsilon_{iass't}, \tag{4}$$

where change in characteristic, $\Delta Q_{ss'}$, equals zero for those addresses which do not experience a change (that is $s = s'$). Our central parameter, β_1, is therefore once again interpreted as an interaction term. This approach yields a number of advantages. Firstly, by including the changes in characteristics directly we encompass the finding above that larger changes are associated with larger responses. This allows us to interpret our results as average marginal effects, which can be used for prediction. Secondly, we can control for other changes in regard to school assignment occurring simultaneously, most notably changes in distance to the district school. We may also employ multiple school characteristics at a time, though we will discuss the somewhat subtle changes in interpretation when multiple measures are included.

Interpreting the results using the specification in (4), we are implicitly assuming symmetric effects, which we have shown may not be present. We however see this as a justifiable simplification for now. When we estimate models of the type specified in (4) we limit our data to SABs where at least one address experiences a change and include only observations two years prior and two years after that change.

We begin by estimating (4) for overall compliance with one school characteristic at a time. The partial results are shown in columns 1-4 in Table 3. Enrollment into the newly assigned school generally increases by around 30 percentage points among those who are actually reassigned, as seen by the value of the parameter on $T \times Post$. The coefficient on distance is significantly negative for all specifications implying that the further a child must travel to the district school the lower the compliance. This is intuitive as travel time likely is associated with a decrease in utility for households.

The interaction terms on SES and NW-share respectively are highly significant. A standard deviation increase of average SES entails an increase in compliance of around $(0.7 \times 0.1) \times 100 \approx 7$ percentage points. Conversely, an increase of ten percentage points in the NW-share decreases the compliance by 4 percentage points. School value added has no discernible effect on compliance.[25]

Interpretation of partial effects Column 5 of Table 3 include all school characteristics in a simple regression. While the coefficient on SVA changes little, we see substantial changes in the parameters on average SES and NW-share. These changes are due to a large negative correlation of 0.84 between

[25]The coefficient on distance, however, doubles when SVA is included, due to the correlation between the official SVA and distance.

	(1)	(2)	(3)	(4)	(5)
T $\times$ Post	0.33***	0.33***	0.33***	0.29***	0.30***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Δ Dist $\times$ Post	-0.03*	-0.03*	-0.03*	-0.08*	-0.06*
	(0.01)	(0.01)	(0.01)	(0.04)	(0.03)
Δ SES $\times$ Post		0.70***			0.58
		(0.14)			(0.37)
Δ NW $\times$ Post			-0.40***		-0.14
			(0.10)		(0.27)
Δ SVA $\times$ Post				0.08	0.09*
				(0.05)	(0.05)
N	53,426	53,426	53,426	48,355	48,355

Table 3: Compliance as a function of SAB change

Columns 1-3 display regression results for the model presented in Equation (4) for one school characteristic at a time and with compliance as dependent variable. Column 4 display the result of an estimation using all characteristics at a time. The models are estimated with "origin-SAB"-year fixed effects. Standard errors are in parentheses and clustered on origin SAB level. † $p < .1$, * $p < .05$, ** $p < 0.01$, ***$p < 0.001$.

	Δ Dist	Δ SES	Δ NW	Δ SVA
Δ Dist	1			
Δ SES	0.0115	1		
Δ NW	0.0815***	-0.836***	1	
Δ SVA	0.0868***	-0.128***	0.189***	1

Table 4: Correlation in school characteristics changes

The table presents the Pearson correlation coefficients for the changes in school measures following a change in SAB weighted by the number of households experiencing the change.† $p < .1$, * $p < .05$, ** $p < 0.01$, ***$p < 0.001$

changes in SES and changes in NW-share as evidenced in Table 4. This correlation makes it extremely difficult to separate out partial effects of changes in socioeconomic and ethnic composition. A subtle issue is whether households can make this distinction themselves. It may be the case, that parents simply use the Non-Western as a proxy for school SES.

As described earlier, we constructed our SES-index from a principal component analysis. This is likely not a perfect measure of SES. If we are willing to assume the NW-share is really just another proxy for (unobserved) socioeconomic composition, we can employ an insight developed by Lubotsky and

Wittenberg (2006) by which we can combine the two estimates to yield a coefficient on a "true" SES-index. If we assume that ΔNW correlates negatively with the unobserved index, then the "true" parameter is given by $0.58 - 0.93 \times (-0.14) = 0.71$, which is almost equal to the parameter value of 0.7 on SES in column 2 of Table 3.[26] This back-of-the-envelope calculation leads us to conclude that the NW-share and our constructed SES index may essentially measure the same underlying socioeconomic conditions and cannot meaningfully be separated in our data.[27] We, therefore, proceed to show results using our SES index and abstain from estimating regressions with more than one school characteristic at a time in what follows.

Margins of response We have previously defined how household may defy the assignment mechanisms through different choices. Because we investigate surprise changes in SABs, we suspect that an important margin is to enroll into school in the original SAB. We therefore further decompose the "other public school"-margin, such that the original district school is mea-

[26]Formally, simplifying notation, we assume that the true SES, SES^*, is approximated by our index, SES, and the Non-Western share, NW, in the following way:

$$\begin{aligned} y &= \beta SES^* + \varepsilon \\ SES^* &= SES + u_1 \\ SES^* &= -\rho NW + u_2, \end{aligned}$$

where we set the coefficient on SES to one and therefore set the scale of the true SES. Then Lubotsky and Wittenberg (2006) show that β may be approximated by simultaneously regressing y on SES and NW and adding up the coefficients according to their covariances as $\beta = \beta_{SES} + \frac{cov(y,NW)}{cov(y,SES)}\beta_{NW}$. With covariance, $cov(y, SES) = 0.000275$ and $cov(y, NW) = -0.000257$ we obtain the values in the main text.

[27]This is naturally a function of the Danish context. The correlations may differ between countries and over time. In the present case, we think of the correlation as being policy-invariant, though this may not be true in the long term. Interestingly, we also find a positive correlation between the average distance change and the change in school value added of 0.09. This correlation most likely reflects that the SVA, as constructed by the Danish Ministry of Education, does not take urbanization into account. We speculate that a high SVA is due to unobserved sorting which correlates with the location decision of households. We see this as problematic for a measure of value added. This issue is reflected in column 4 of Table 3 where the coefficient on distance drops to zero when we include SVA. We however have more confidence in our distance measure than the SVA provided by the ministry.

sured separately. We therefore modify equation (1) to encompass all responses:

$$comply = 1 - third - original - priv - move, \tag{5}$$

where $comply$ takes the value of one if the child enrolls in the assigned district school, $original$ denotes the departure school and $third$ denotes a public school different from original school and the newly assigned school. If the family moved out of the district between age 5 and 7 of the child, $move$ takes the value of one.

We decompose non-compliance rate by estimating models corresponding to column 2 in Table 3 using each component from equation (2). The results are displayed in Table 5. Column 1 reproduce column 2 of Table 3. The remaining columns approximately sum to (minus) the first column. It is evident that changes in compliance stem from the publicly provided option of choosing other public schools, measured by the outcomes, “Original” and “Third”. When school-SES increases, the majority of the increase in compliance stems from a diminishing propensity to choose other public schools, as seen by the estimate of -0.79 on SES in the column denoted by “Third” in Table 5. An additional source of increased compliance is the decrease in private school enrollment, though this effect is only significant at a ten-percent confidence-level. This increase in compliance is however attenuated by an increase in the propensity to stay in the original district school. These results are consistent with our first difference-in-difference analysis where we categorized shocks as positive and negative, see Figure B.1. We show below that this response along the $original$-margin stems mostly from low SES households staying behind.

Household heterogeneity in responses - SES As mentioned in section 3 preferences and constraint likely differ at the household-level. This would lead to heterogeneity in responses to changes in SABs. To elicit this, we interact the model presented in Equation (4) with household-level charac-

	Comply	Original	Third	Private	Move
T × Post	0.33***	-0.13***	-0.18***	-0.03*	0.01
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)
Δ Dist × Post	-0.03*	0.01	0.04***	-0.02†	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Δ SES × Post	0.70***	0.38*	-0.79***	-0.19†	-0.10
	(0.14)	(0.15)	(0.12)	(0.11)	(0.09)
N	53,426	53,426	53,426	53,426	53,426

Table 5: Responses along different margins

Columns display regression results for the model presented in Equation (4) using SES as a measure of schools with different dependent variables, displayed in the columns title. The models are estimated with "origin-district"-year fixed effects. Standard errors are in parentheses and clustered on origin district level. † $p < .1$, * $p < .05$, ** $p < 0.01$, ***$p < 0.001$

teristics. We begin by fully interacting the model with SES quartile of the household. The effect on overall compliance along with different margins of response is shown in Table 6. The basic reaction to an attendance boundary change along with distance changes exhibits no heterogeneity, as evidenced by the first eight rows. The results, however, indicate a high degree of heterogeneity in responses to changes in school SES. In column 1 the coefficient on average SES is monotonically increasing in own SES. In other words, the higher the socioeconomic status of the household, the larger the expected response. The response from a change in school SES is around 2.5 times larger for the highest quartile compared to the lowest quartile.[28].

The low effect for households in the first SES quartile is explained by a large tendency to stay behind in the old SAB, a choice higher SES-households are unlikely to make, being much more likely to comply with the assignment when average SES increases. We find weak and insignificant evidence that private school may also be a margin of response which higher SES households exploit (though not the highest SES quartile.)

To give a sense of the overall magnitude of changes in cohort size and average SES following an attendance boundary change, we perform a back-

[28]Calculated as $\frac{0.53+0.36}{0.36} \approx 2.48$

	Comply	Original	Third	Private	Move
T × Post	0.33***	-0.09*	-0.20***	-0.03	-0.01
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
T × Post × SE Q2	-0.01	-0.02	0.02	-0.02	0.03
	(0.02)	(0.04)	(0.03)	(0.03)	(0.03)
T × Post × SE Q3	-0.01	-0.07†	0.02	0.03	0.03
	(0.02)	(0.04)	(0.03)	(0.02)	(0.02)
T × Post × SE Q4	-0.00	-0.04	-0.00	0.03	0.02
	(0.03)	(0.04)	(0.03)	(0.03)	(0.03)
Δ Dist × Post	-0.03**	-0.01	0.04***	-0.01	0.01
	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Δ Dist × Post × SE Q2	0.00	0.01	-0.00	-0.00	-0.01
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Δ Dist × Post × SE Q3	0.00	0.02*	0.00	-0.01†	-0.01
	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)
Δ Dist × Post × SE Q4	-0.01	0.02	-0.00	-0.01	-0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Δ SES × Post	0.36*	0.65**	-0.67**	-0.16	-0.18
	(0.15)	(0.24)	(0.25)	(0.17)	(0.14)
Δ SES × Post × SE Q2	0.23†	-0.34	0.00	-0.02	0.12
	(0.13)	(0.25)	(0.19)	(0.20)	(0.17)
Δ SES × Post × SE Q3	0.37**	-0.29	-0.21	-0.03	0.16
	(0.14)	(0.23)	(0.27)	(0.19)	(0.18)
Δ SES × Post × SE Q4	0.53**	-0.29	-0.29	-0.01	0.06
	(0.20)	(0.27)	(0.28)	(0.20)	(0.16)
N	47,498	47,498	47,498	47,498	47,498

Table 6: Heterogeneity in responses along different margins, interacted with househols SES

Columns display regression results for the model presented in Equation (4) using SES as a measure of schools with different dependent variables, displayed in the columns title. Characteristics are interacted with household SES quartile. The first quartile is baseline.The models are estimated with “origin-SAB”-year fixed effects. Standard errors are in parentheses and clustered on origin SAB level. † $p < .1$, $* \ p < .05$, $** \ p < 0.01$, $* * * p < 0.001$

of-the-envelope calculation using the results in Table 6. We assume that the group of children to be transferred has a measure of one and is uniformly distributed in four quartiles on the unit interval along the SES dimension.[29] First, assume that this group is transferred to a new school but experience no change in school SES, formally $\Delta SES = 0$. The group will have a mass of around 0.33 and an SES of 0.5 equal to the reference group. If school SES falls by 1 std. the mass falls further to 0.27, i.e. a decrease in compliance rate by 20 pct. Not only does the mass fall but the average SES of enrolling households are now 0.48, i.e. a drop of 4 pct.[30] Although changes in SES have a linear effect on compliance rate the impact on SES is non-linear. A drop in school SES of resp. 2 and 3 std. entail a drop in SES of complying households to resp. 0.45 and 0.38, i.e. a drop of resp. 11 and 25 pct. This implies that not only does the size of the group fall, but the socioeconomic composition will be markedly different from the group which was intended to enroll in the new school. Both of these factors should be taken into account when policy-makers are considering redrawing attendance boundaries.

Household heterogeneity in responses - Ethnicity We round off the analysis of attendance boundary changes by investigating heterogeneity in responses across ethnic groups. We interact our basic model with an indicator for whether the child in the household is of Non-Western descent. The results of this exercise are displayed in Table 7. Compared to the basic estimation in column 3 of Table 3, the parameter of the reference group changes from -0.4 to -0.55, which imply that Western/Danish households react more to changes in the Non-Western share than average. Conversely, Non-Western households react to a far lesser degree, as can be seen by adding up the appropriate coefficients in the “Comply”-column (-0.55 + 0.49 = -0.06).

[29] For each quartile we assign an SES in the middle of the interval.
[30] This is calculated using Equation 10.

	Comply	Original	Third	Private	Move
T × Post	0.34***	-0.15***	-0.18***	-0.02	0.01
	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)
Δ Dist × Post	-0.03*	0.01	0.04***	-0.02†	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Δ NW × Post	-0.55***	-0.18†	0.61***	-0.03	0.16†
	(0.12)	(0.10)	(0.09)	(0.09)	(0.09)
T × Post × NW	-0.02	0.10**	-0.02	-0.05†	-0.01
	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Dist × Post × NW	0.00	-0.05*	0.01	0.02†	0.02
	(0.01)	(0.02)	(0.01)	(0.01)	(0.02)
Δ NW × Post × NW	0.49***	-0.26†	-0.17	0.06	-0.11
	(0.13)	(0.15)	(0.16)	(0.11)	(0.12)
N	53,426	53,426	53,426	53,426	53,426

Table 7: Heterogeneity in responses along different margins, interacted with households' SES

Columns display regression results for the model presented in Equation (4) using SES as a measure of schools with different dependent variables, displayed in the columns title. Characteristics are interacted with household Non-Western dummy. The models are estimated with "origin-SAB"-year fixed effects. Standard errors are in parentheses and clustered on origin SAB level. † $p < .1$, $* \, p < .05$, $** \, p < 0.01$, $* * * p < 0.001$

Concluding attendance boundary changes

We have documented that household compliance to school assignment is a function of the school characteristics, with a stark social gradient. A consistent finding is that other public schools make out the primary way by which households avoid the reassignment to a new school. The most important means of non-compliance is therefore publicly provided.

We find very small effects of changes on private school enrollment. We conjecture that the lack of response along this margin may be due to the surprise effect of the redistricting. If households have not foreseen the change, as we indeed assume, then a private school may not be an option due to oversubscription. Thus, responses along this margin is probably larger in the long run, which we cannot capture using the changes in attendance boundaries. The same kind of reasoning may apply to the decision to move out of the district, which may entail significant costs. To investigate whether results may differ in the long term we turn to our auxiliary approach.

6 Auxiliary approach: Cross border comparisons

Our aim in this section is to the uncover responses, which may not have been feasible for households during the short window used to measure the impact of district changes in the previous section. We once again investigate all three options for opting-out of the school district (i.e. non-compliance); relocating, enrolling in either private and exploiting the public loophole by enrolling in another public school.

As noted in the previous section, the school-SES and the NW-share are highly correlated in a Danish context. In this setup, we observe a correlation of school borders' differences in average socioeconomic status and share with a non-Western heritage of -0.75, see Appendix Table C1. We, therefore, focus on socioeconomic status and note that replacing SES with NW-share will yield the same conclusions.

We begin our analysis with reduced form estimation of border differences without covariates. We follow the approach presented by Bayer et al. (2007) and construct bins of distances. We define a distance to be negative if the household is associated with the school with the lowest value of the two neighbor schools on a given school characteristic. We implement regressions of the following kind:

$$Y_{ibt} = \sum_{d=D^-}^{D^+} \gamma^d \mathbf{1}(dist_{ibt} = d) + \mu_{bt} + \varepsilon_{ibt}, \tag{6}$$

where Y_{ibt} is our binary dependent variable for child in household i at border b at time t, $dist$ is the signed distance to the border and μ_{bt} is a border-year fixed effect. By including the border-year-fixed effect we are comparing only within border regions in the same year. As we are interested in the difference across borders, we center our results on a left side dummy (i.e. the lower side).[31]

[31]We include a dummy for being on the higher value border side as well as interactions of high side with distance. The low side is therefore the reference category.

To provide central estimates to compare with the previous section we collapse Equation (6) to the following:

$$Y_{ibt} = \gamma^{+}\mathbf{1}(dist_{ibt} > 0) + \mu_{bt} + \varepsilon_{ibt}, \tag{7}$$

where γ^{+} denotes the value of being on the high side of the boundary. We estimate the marginal effect of school characteristics on compliance as equation (7), but replacing the dummy for being o a high side with the SES of the district school. Due to the inclusion of border-year-fixed effects the only variation used is from crossing the border. We report these estimates in the figures as well.

In our samples, we use all children age 5 located residing less than 2 kilometers from a border. Each child may be located within multiple border regions. We cluster standard errors to accommodate this. Outcomes are observed at age 7, ie. where all children should be enrolled in a primary school.

Socioeconomic status We estimate Equation (6) for the probability of choosing a non-district public school, private school and moving away, again using school average SES as a measure of schools. The resulting γ^{d} from the estimations are plotted in Figure 7.[32] We start by noting that the probability of non-compliance increases as one approaches the border from either side, as seen in Figure 7a. This can be ascribed to the decreased distance to the neighboring school closer to the border.[33] Despite this pattern there is

[32]Appendix Figure C.2 presents estimates for all three measures, SES, NW-share and SVA (as provided by the ministry).

[33]We note that the rising tendency to not attend the district school closer to the border is fully explained by enrollment in other public schools and moving with no such pattern for private schools. Therefore, the pattern is likely the product of commuting distances. As one approaches the border the distance to the neighbor school will decrease. It may, therefore, be more *convenient* for households to choose the neighbor school. If school characteristics are not important we should thus observe a bell shape around the border. The same logic can account for the rise in movement propensity when approaching the border from either side in Figure 7d.

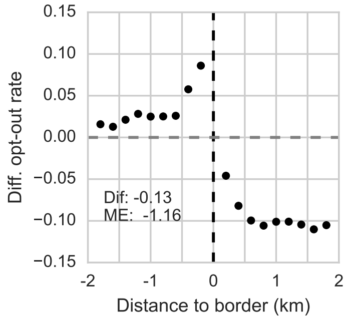
a clear discontinuity at the border taking the form of a vertical shift in non-compliance. In other words, close to the border of two attendance zones, being on the side with lower average student SES is associated with around 13 percentage points lager non-compliance. We interpret this as evidence that socioeconomic compositions of the student body matter for household compliance rates.

Once again, we see that the public option is the main source of non-compliance. Figure 7b shows that choosing another public school makes out over half the total non-compliance. This is in line with the Difference-in-Difference results. In the previous section, we found no clear evidence that households use private school as a means by which to defy the assignment mechanism. In Figure 7c, however, we see that differences in private school enrollment make out a sizable portion of the total difference in compliance rate. The difference in private school enrollment of around three percentage points accounts for a quarter of the total non-compliance. An equal portion of the non-compliance is explained by households moving out of the district before the child turns seven, as seen in Figure 7d.

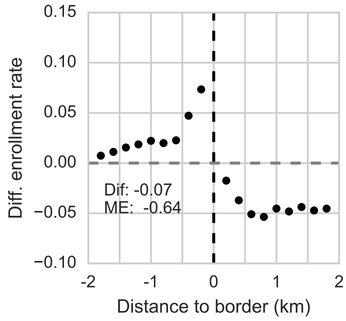
When we rescale the average difference in non-compliance with the average difference in school SES the estimate is -1.16 which is numerically higher than our baseline estimate of -0.7 from our Difference-in-Difference approach, see Table 3. However, subtracting the effect from private school and moving, one gets the public option exclusively, whit an estimate of -0.64 - much closer to the effect identified off the changes in attendance boundaries. This implies that the long-term estimates using the Boundary Discontinuity Design are in line with the Difference-in-Difference estimate though the former is subjective to more restrictive identifying assumptions.[34]

The results displayed in Figure 7 are averages over all borders. However, an issue is that differences between the neighboring district schools are of different magnitudes. If the school quality measures actually cause the

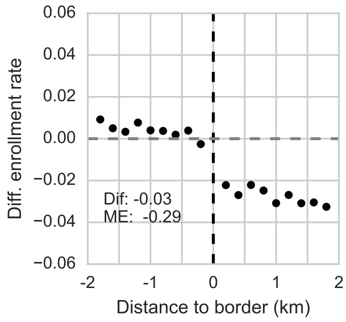
[34]We perform a number of robustness tests including adjusting for covariates and exploit heterogeneous effects. The results are robust to a variety of specifications.



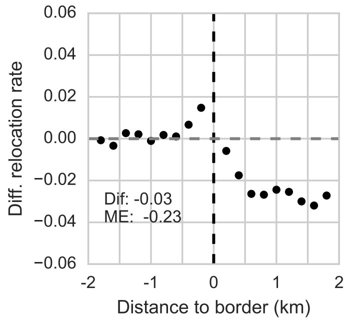
(a) Total non-compliance



(b) Other public school



(c) Private school



(d) Move before age 7

Figure 7: Boundary difference in opt-out for low/high SES schools

The figures depict the estimated parameters of a BDD-model estimated for different dependent variables and average SES as school characteristic according to the model presented in equation (6). The dependent variable in Figure 7b is a dummy which takes the value of one if the child is enrolled in a non-district public school school. The dependent variable in Figure 7c is a dummy indicating that the child is enrolled a private school school. The dependent variable in Figure 7d is a dummy indicating that the household has moved before before the child turns seven years old. Negative distance to border signifies that the household is situated in the district of the two bordering districts with the lower value of the school characteristic. The models are estimated with fixed effects at the border-year level. The mean difference is estimated in OLS and displayed in the lower-left corner of the figures. The corresponding rescaled estimate, denoted ME, is estimated with OLS by replacing the indicator for being on the right side with the average SES of schools on either side. All marginal effect estimates are significant at $p < 0.001$; see Table C2 for details.

observed differences, we would expect the differences in opting-out use to be monotonically increasing in the size of differences. To investigate this, we estimate the model from Equation (7) across the distribution of border-SES-differences for opting out, decomposed into the three elements; another public school, private school and moving away. The results of this exercise are displayed in Appendix Figure C.3. The decomposition show that all three sources of non-compliance generally are important and that differences in behavior are almost linear in differences in school characteristics, which imply constant marginal effects. [35]

Response heterogeneity by socioeconomic status We finish our boundary discontinuity analysis of the importance of school SES by investigating heterogeneity in responses. We repeat the analysis for each quartile of the household SES-distribution.[36] The results of this exercise is presented in Figure 8. Figure 8a shows a clear discontinuity in compliance for all quartiles. However, the implied marginal effect of the highest quartile, -1.47, is more than twice the size of the estimate for the lowest quartile of -0.67. Thus, high SES households are much more sensitive to the socioeconomic composition in the district school when choosing whether to comply with the school assignment. Figure 8b shows the difference in enrollment into non-district public schools. A clear discontinuity for all quartiles is evident, but all exhibit marginal effects around -0.5. The heterogeneity is therefore not explained by the public school option in this setting. Figure 8c shows private schools enrollment is more heterogeneous across SES quartiles - the discontinuity is monotonically increasing in household SES. The ratio of the marginal effect of the highest to the lowest quartile is almost 5.[37] In other words, faced

[35]Note the close resemblance to the Wald-estimator. Again, subject to an exclusion restriction the discontinuity estimates may be rescaled by the difference in school SES to obtain an IV estimate of the marginal effect of average school SES on compliance rates. The very linear relationship in Appendix Figure C.3 implies a constant marginal effect.

[36]This approach, therefore, compares *within* SES quartile *within* border-year and is equivalent to fully interacting the border-year fixed effect with SES-quartile.

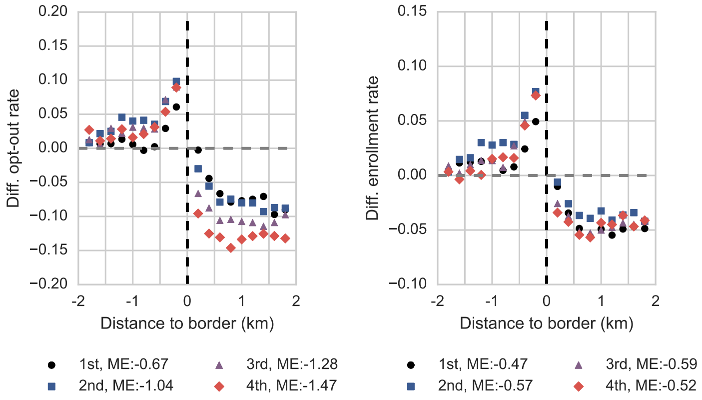
[37]From Figure 8c we calculate the ratio $(-0.69)/(-0.14) = 4.9$

with lower average SES in the district school a high-SES household is much more likely to enroll the child in private school than a low-SES household. Figure 8d shows the same pattern for the relocation of households as for private school enrollment, though the ratio of highest to lowest quartile is only 3.4. In other words, high SES-households are much more likely to move away when faced with a low-SES district school, though they are relatively more sensitive along the private school margin compared to low-SES households. In this framework we cannot investigate whether this socioeconomic gradient is due to preferences or constraints, but we conjecture that both likely play a role. Regardless of the source of heterogeneity in behavior, our findings imply that redrawing of attendance boundaries likely will lead to less between school homogeneity than a prediction without consideration of behavioral effects would suggest.

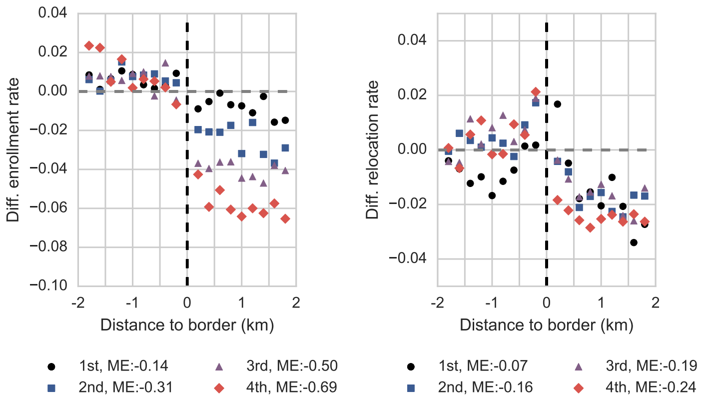
School value-added We investigate the importance of school value-added in Appendix C. Figure C.2 shows the discontinuities in responses to differences in school value-added; the responses are weaker than the responses for SES in Figure 7. An overview of estimated marginal effects associated with the differences is found in Table C3. Figure C.4 shows decomposition of non-compliance as a function of school value-added; it is evident that increased school value-added is associated with higher compliance, though the effects are less clear than those for other school characteristics.

7 Conclusion

Policy-makers who aim to balance school composition can manipulate the school boundaries and choose which children are supposed to enroll where. But the efficacy of this strategy, as with most other public interventions, is threatened by individuals' behavioral responses. We have documented that parents react to redistricting by opting out of their assigned school. They do this by moving, choosing other public schools and private schools. Households with high socioeconomic status drive the responses, which implies that



(a) Total non-compliance (b) Other public school



(c) Private school (d) Move before age 7

Figure 8: Boundary difference in opt-out for low/high SES schools - by household SES

The figures depict the estimated parameters of a BDD-model estimated for different dependent variables and average SES as school characteristic according to the model presented in equation (6). The model is estimated separately for each quartile of household SES. The dependent variable in Figure 8b is a dummy which takes the value of one if the child is enrolled in another public school. The dependent variable in Figure 8c is a dummy indicating that the child is enrolled in a private school. Negative distance to border signifies that the household is situated in the district of the two bordering districts with the lower value of the school characteristic. The models are estimated with fixed effects at the border-year level. The mean difference is estimated in OLS and displayed in the lower-left corner of the figures. The corresponding rescaled estimate is estimated with OLS by replacing the indicator for being on the right side with the average SES of schools on either side. For parameter coefficients and tests from the estimation of marginal effects see Table C2.

the “leakage” occurs in the top of the distribution. Consequently, if policy-makers want to minimize variance in student compositions they must do so under the constraint that parents have an outside option.

Interpreted more generally these findings imply that there are limits to the possible manipulation of peer groups when individuals have an outside option. In other words, there is limited potential for what Durlauf (1996a) refer to as associational redistribution. Consequently, our results matter not only for drawing boundaries between schools. They are also relevant in a broader sense for designing groups within schools and organizations.

Our results indicate some possible venues for further analysis. Our model is limited by only investigating enrollment decisions and relocation partially; it would be interesting to model the choice of residence and school simultaneously. One possible theoretical analysis would be to investigate the extent to which limiting the outside options of households for enrollment in private schools affects the compliance of parents in the public district school system.

A peculiar facet of the Danish system is the high (but not full) degree of public financing of private schools and the possibility of enrolling in non-district public schools. The latter is in practice a very opaque process. It is likely that this process makes it relatively easier for highly sophisticated (and most likely educated) parents to exploit the system to the detriment of less sophisticated parents. The private school funding and the implementation of choice mechanisms make it possible for households to segregate in the educational system without the usually associated residential segregation. This may be beneficial if there are benefits to desegregation outside primary school education. Nevertheless, the implied decoupling between educational and residential segregation also diminishes the efficacy of redistricting as a policy tool to increase equality of opportunity.

References

Abdulkadiroğlu, A., Pathak, P., Schellenberg, J., Walters, C. R. 2017. Do parents value school effectiveness?, Working paper.

Abdulkadiroğlu, A., Sönmez, T. 2003. School Choice: A Mechanism Design Approach. American Economic Review, 93, 729–747.

Baum-Snow, N., Lutz, B. F. 2011. School desegregation, school choice, and changes in residential location patterns by race. American Economic Review, 101, 3019–46.

Bayer, P., Ferreira, F., McMillan, R. 2007. A Unified Framework for Measuring Preferences for Schools and Neighborhoods. Journal of Political Economy, 115, 588–638.

Bifulco, R., Ladd, H. F., Ross, S. L. 2009. Public school choice and integration evidence from durham, north carolina. Social Science Research, 38, 71–85.

Bjerre-Nielsen, A., Gandil, M. H. 2018b. Privacy in spatial data with high resolution and time invariance.

Bjerre-Nielsen, A., Gandil, M. H. 2018a. The price of free schools.

Black, S. E. 1999. Do Better Schools Matter? Parental Valuation of Elementary Education. The Quarterly Journal of Economics, 114, 577–599.

Black, S. E., Devereux, P. J. 2011. Recent developments in intergenerational mobility. 4, Elsevier B.V. 1487–1541.

Borghans, L., Golsteyn, B. H. H., Zölitz, U. 2015. Parental preferences for primary school characteristics. B.E. Journal of Economic Analysis and Policy, 15, 85–117.

Burgess, S., Greaves, E., Vignoles, A., Wilson, D. 2011. Parental choice of primary school in England: what types of school do different types of family really have available to them?. Policy Studies, 32, 531–547.

Burgess, S., Greaves, E., Vignoles, A., Wilson, D. 2015. What Parents Want: School Preferences and School Choice. The Economic Journal, 125, 1262–1289.

Chetty, R., Hendren, N., Katz, L. F. 2016. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. American Economic Review, 106, 855–902.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., York, R. L. 1966. Equality of educational opportunity..Technical report, U.S. Department Of Health, Education Welfare, Office of Education, Report Number 0E-36001.

Durlauf, S. N. 1996a. A theory of persistent income inequality. Journal of Economic Growth, 1, 75–93.

Durlauf, S. N. 1996b. Associational Redistribution: A Defense. Politics & Society, 24, 391–410.

Fack, G., Grenet, J. 2010. When do better schools raise housing prices? Evidence from Paris public and private schools. Journal of Public Economics, 94, 59–77.

Gibbons, S., Machin, S., Silva, O. 2013. Valuing school quality using boundary discontinuities. Journal of Urban Economics, 75, 15–28.

Gillies, S. et al. 2007–. Shapely: manipulation and analysis of geometric objects.

Hagberg, A. a., Schult, D. a., Swart, P. J. 2008. Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008), 836, 11—-15.

Hastings, J. S., Kane, T. J., Staiger, D. O. 2009. Heterogeneous preferences and the efficacy of public school choice. Working paper.

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. Computing in Science and Engineering, 9, 99–104.

Imberman, S. A., Lovenheim, M. F. 2016. Does the market value value-added? evidence from housing prices after a public release of school and teacher value-added. Journal of Urban Economics, 91, 104–121.

Jones, E., Oliphant, T., Peterson, P. et al. 2001–. SciPy: Open source scientific tools for Python. [Online; accessed Jan. 10, 2018].

Lubotsky, D., Wittenberg, M. 2006. Interpretation of regressions with multiple proxies. The Review of Economics and Statistics, 88, 549–562.

McKinney, W. 2010. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 1697900, 51–56.

Monarrez, T. 2017. Attendance boundary policy and the segregation of public schools in the united states. Working paper.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É. 2012. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Rangvid, B. S. 2009. School choice, universal vouchers and native flight from local schools. European Sociological Review, 26, 319–335.

Riedel, A., Schneider, K., Schuchart, C., Weishaupt, H. 2010. School choice in german primary schools: How binding are school districts? 1/schulwahl in deutschen grundschulen: Wie verbindlich sind schulbezirke? Journal for educational research online, 2, p. 94.

Rossell, C. H. 1975. School desegregation and white flight. Political Science Quarterly, 90, 675–695.

Sacerdote, B. 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far? In Handbook of the Economics of Education, 3, Elsevier, 249–277.

Saporito, S., Sohoni, D. 2007. Mapping educational inequality: Concentrations of poverty among poor and minority students in public schools. Social Forces, 85, 1227–1253.

Seabold, S., Perktold, J. 2010. Statsmodels: econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, 57–61.

Yinger, J. 1986. Measuring racial discrimination with fair housing audits: Caught in the act. The American Economic Review, 881–893.

A Data description

This appendix consists of an account for how we measure socioeconomic status and also contains additional descriptive output.

A.1 Construction of SES index

This sub-appendix outlines how we construct our socio-economic index. We describe our approach of reducing a set of socio-economic variables to a single socio-economic index (SES index henceforth) and we evaluate the index' performance.

We construct our SES-index by choosing the first variable resulting from a principal component analysis (PCA) based on the following variables:

- INC: We calculate the market income rank of all adults in the population. We select the highest income rank observed in a household.
- LCE: A dummy which takes the value of one if an adult in a household has completed a long cycle education.
- NE: A dummy which takes the value of one if an adult in a household has not completed in education beyond primary school or have no registered education.
- EMP: A dummy which takes the value of one if an adult in a household is employed.

We select the first component of the PCA. This leads to the following index:

$$SES = 0.62INC + 0.38LCE - 0.44NE + 0.53EMP, \tag{8}$$

where all variables have been standardized to their corresponding z-scores. This index accounts for 47 pct. of the variation in the four variables. The SES-index applied in our paper is the population ranks of SES, as such it is uniformly distributed on the unit interval.

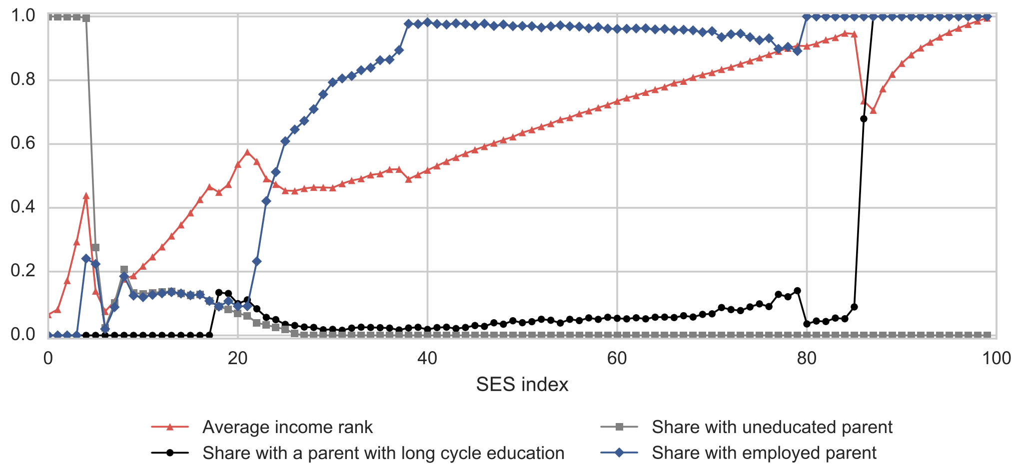
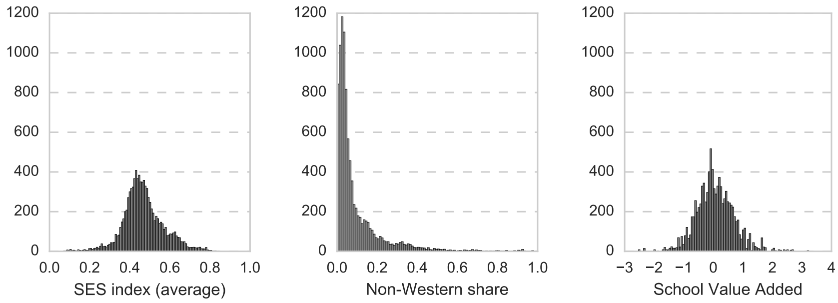


Figure A.1: Average characteristics as a function of SES-index

The figure depicts means of variables used to construct the SES-index. The SES-index is uniformly distributed on the unit interval. Each marker represent the mean of the variable in question within a percentile bin. Income rank is bounded between 0 and 1.

To get a sense of the mapping between the underlying variables we calculate averages of the underlying variables in percentiles of the SES-index. The results are displayed in figure A.1. While this is a very simple index we find that this component is intuitive. In the bottom of the distribution almost all households have an uneducated parent and no parent with a high cycle education. In the top 75 percent of the distribution no household contain an uneducated parent. Income and employment are both rising in the SES-index. Thus we find it safe to assume that the SES-index reflects a true underlying socioeconomic status.



(a) Average SES (b) NW-share (c) SVA

Figure A.2: Distributions of school characteristics

The figures depict the distributions of school characteristics for district schools for the years 2008-2015. Figures A.2a, A.2b and A.2c respectively display district schools' average share of non-Western descendants, average SES-index and school value added. Note these measures exclude private schools.

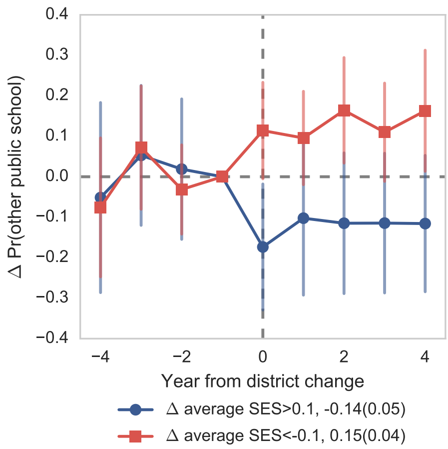
B Supplementary results for changes in district borders

In this appendix we provide additional results for the main approach. We begin with the computation of SES for complying households.

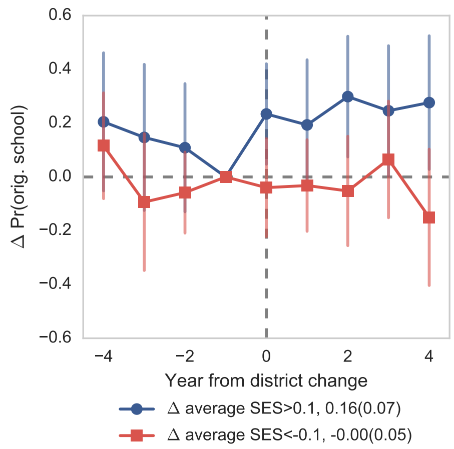
$$E[SES_i|comply_{i,ss'} = 1,\, \Delta SES] = \frac{\sum_{q\in\{1,..,4\}} w_q \left(\beta_q^{P\cdot T} + \beta_q^{P\cdot T\cdot SES}\Delta SES_{ss'}\right)\cdot \mu_q}{\sum_{q\in\{1,..,4\}} w_q \left(\beta_q^{P\cdot T} + \beta_q^{P\cdot T\cdot SES}\Delta SES\right)} \tag{9}$$

where q denotes quartile, ΔSES is the change in school SES and μ_q is the mean SES for quartile q. We can rewrite the equation into and plug in parameter estimates from Table 6 into Equation 9:

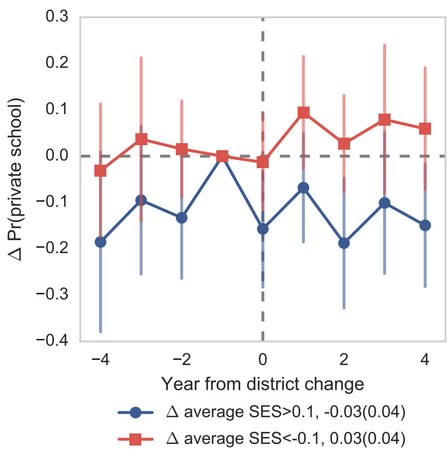
$$E[SES_i|comply_{i,ss'} = 1,\, \Delta SES_{ss'}] = \frac{0.165 + 0.388\cdot \Delta SES}{0.329 + 0.663\cdot \Delta SES} \tag{10}$$



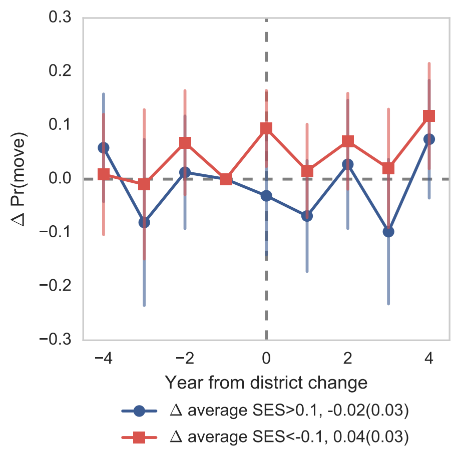
(a) Other public school



(b) Stay in original district school



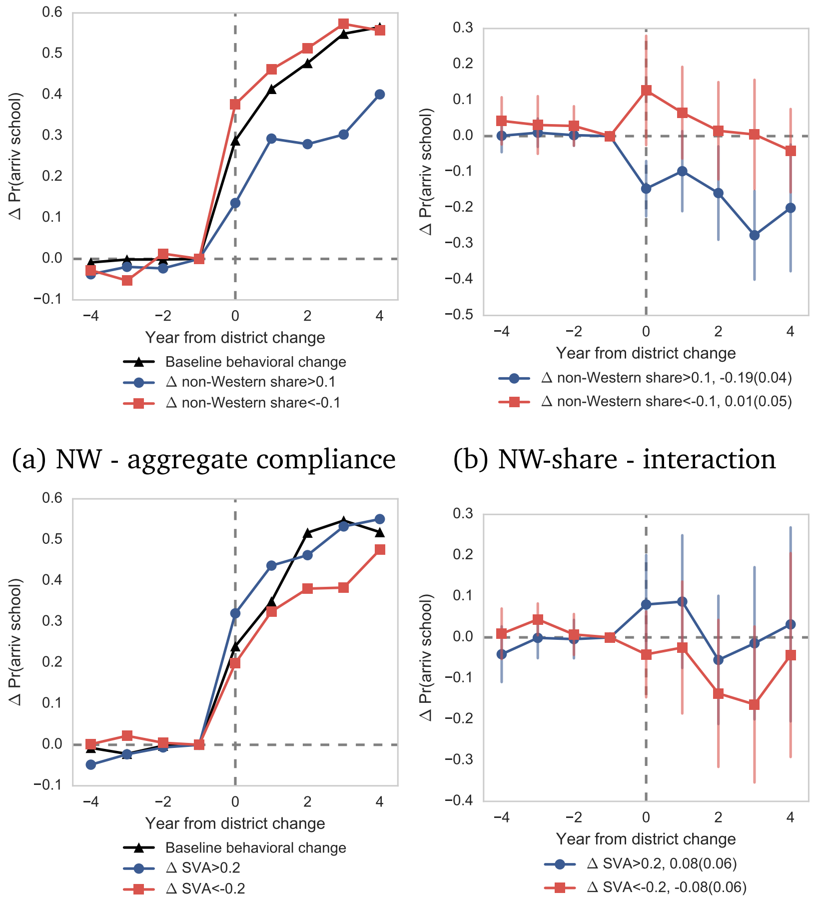
(c) Private school enrollment



(d) Moving

Figure B.1: Response along margins to change in district by school characteristic

The figure display the interaction terms, β_{-}^{k} and β_{+}^{k}, along with 95-percent confidence intervals. The parameters represent the difference in likelihood of enrolling in the new district school when the average SES at a school level changes relative to the average arrival probability following a district change. The dependent variables of all figures are binary and measured at age 7 based on the district at age 7 for address at age 5. The models are estimated with "origin-district"-year fixed effects. Standard errors are clustered on origin district level. Results are centered at the year before the district change. Estimates from a simple before-after-DID are reported in the legends of figure 5b.



(a) NW - aggregate compliance (b) NW-share - interaction

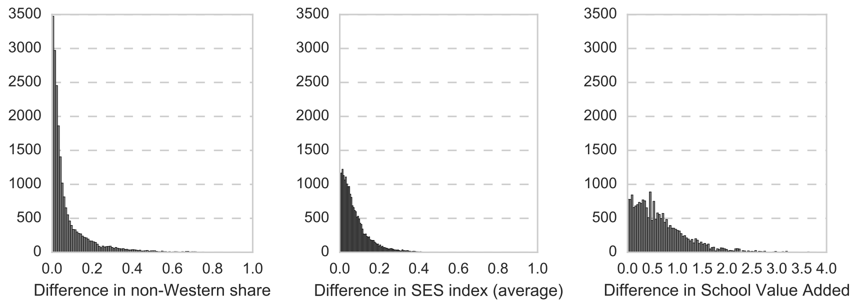
(c) SVA - aggregate compliance (d) SVA - interaction

Figure B.2: Compliance as a function of Non-Western share and school value-added

The figures on the left display changes in estimated compliance rates based on the model in (2) estimated with differnent measures of school characteristics. The black lines depict the estimated β_T^ks, while the blue and red line depict $\beta_T^k + \beta_-^k$ and $\beta_T^k + \beta_+^k$ respectively. The figures to the right display the interaction terms, β_-^k and β_+^k, along with 95-percent confidence intervals. The parameters represent the difference in likelihood of enrolling in the new district school when the school characteristic at a school level changes relative to the average arrival probability following a district change. The dependent variable is binary and equals one if the child is enrolled in the district school at age 7 based on the district at age 7 for address at age 5. The y-axis denotes the excess probability of enrolling relative to baseline. Standard errors are clustered on origin district level. Results are centered at the year before the district change. Estimates from a simple before-after-DID are reported in the legends of figure 5b.

C Supplementary results for auxiliary approach: Cross border comparison

This appendix provides supporting information with additional results for the analysis using border comparisons in Section 6.



(a) NW-share (b) Average SES (c) SVA

Figure C.1: Border differences of school characteristics

The figures depict the distributions of boundary differences in school characteristics for the years 2008-2015. Figures C.1a, C.1b and C.1c display the associated absolute differences in the school measures between neighboring districts.

	ΔSES_b	ΔEMP_b	ΔINC_b	ΔHCU_b	ΔNE_b	ΔNW_b	ΔGPA_b	ΔSVA_b
ΔSES_b	1.00	0.88	0.98	0.74	-0.83	-0.75	0.66	0.18
ΔEMP_b	0.88	1.00	0.89	0.45	-0.83	-0.84	0.58	0.15
ΔINC_b	0.98	0.89	1.00	0.66	-0.83	-0.80	0.64	0.17
ΔHCU_b	0.74	0.45	0.66	1.00	-0.45	-0.39	0.56	0.14
ΔNE_b	-0.83	-0.83	-0.83	-0.45	1.00	0.76	-0.58	-0.16
ΔNW_b	-0.75	-0.84	-0.80	-0.39	0.76	1.00	-0.51	-0.10
ΔGPA_b	0.66	0.58	0.64	0.56	-0.58	-0.51	1.00	0.23
ΔSVA_b	0.18	0.15	0.17	0.14	-0.16	-0.10	0.23	1.00

Table C1: Correlation matrix for border school district differences across borders

The table presents a correlation matrix for variables used in the analysis and the variables used to construct the socioeconomic index.

	Any	Q1	Q2	Q3	Q4
Opt-out of district	-1.159^{***}	-0.674^{***}	-1.038^{***}	-1.279^{***}	-1.466^{***}
	(0.032)	(0.041)	(0.044)	(0.047)	(0.046)
Non-district public school	-0.641^{***}	-0.466^{***}	-0.567^{***}	-0.585^{***}	-0.518^{***}
	(0.029)	(0.040)	(0.033)	(0.036)	(0.036)
Private school	-0.287^{***}	-0.141^{***}	-0.305^{***}	-0.495^{***}	-0.686^{***}
	(0.014)	(0.024)	(0.024)	(0.028)	(0.033)
Relocation to new district	-0.226^{***}	-0.073^{**}	-0.164^{***}	-0.190^{***}	-0.243^{***}
	(0.019)	(0.027)	(0.025)	(0.021)	(0.024)

†: $p < .1$, $*$: $p < 0.05$, $**$: $p < 0.01$, $***$: $p < 0.001$

Table C2: IV estimates of difference in opt-out rate from difference in socioeconomic status

The table presents estimations of the marginal propensity to opt out as a function of difference in school average socioeconomic status (SES). School SES is instrumented by a dummy for being at the "high side" of the district border. Models are estimated with only one quality measure at a time. Standard errors are clustered by the connected component of a graph where an edge exists between border regions if the same household forms part both regions. This procedure creates in 480 clusters.

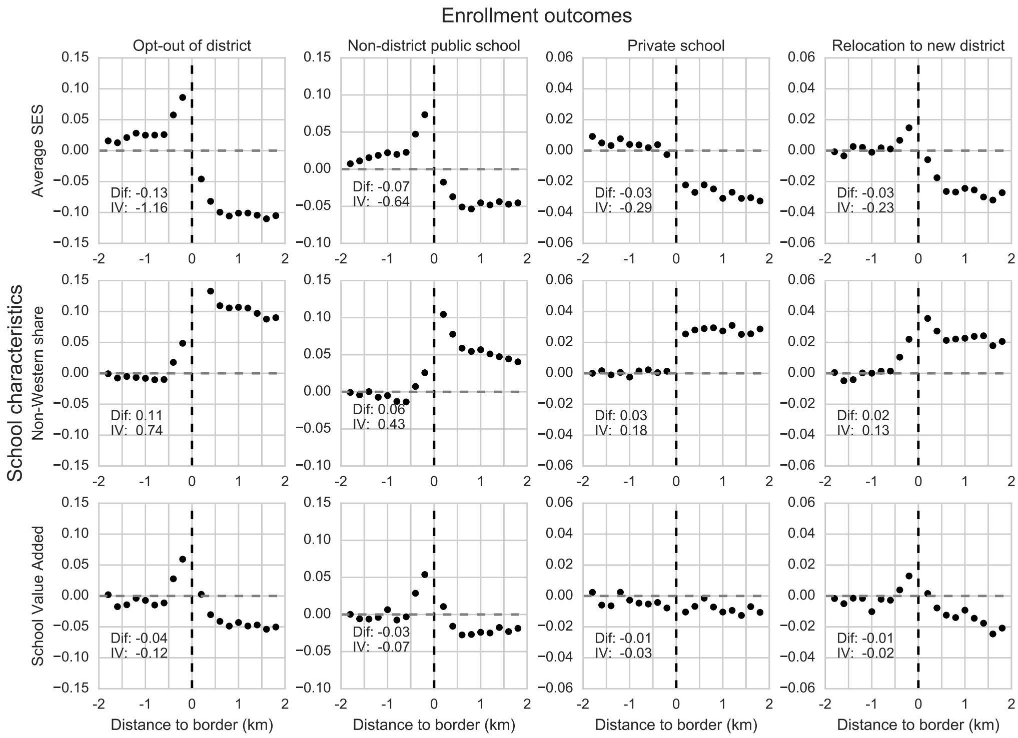


Figure C.2: Differences in opt-out rate as a function of difference in school characteristics

The figures depict the estimated parameters of a BDD-model estimated for different dependent variables and School Value Added as school characteristic according to the model presented in equation (6). The dependent variable in Figure 7b is a dummy which takes the value of one if the child is enrolled in a non-district public school school. The dependent variable in Figure 7c is a dummy indicating that the child is enrolled a private school school. The dependent variable in Figure 7d is a dummy indicating that the household has moved before before the child turns seven years old. Negative distance to border signifies that the household is situated in the district of the two bordering districts with the lower value of the school characteristic. The models are estimated with fixed effects at the border-year level. The mean difference is estimated in OLS and displayed in the lower-left corner of the figures. The corresponding rescaled estimate is estimated with OLS by replacing the indicator for being on the right side with the average SES of schools on either side.

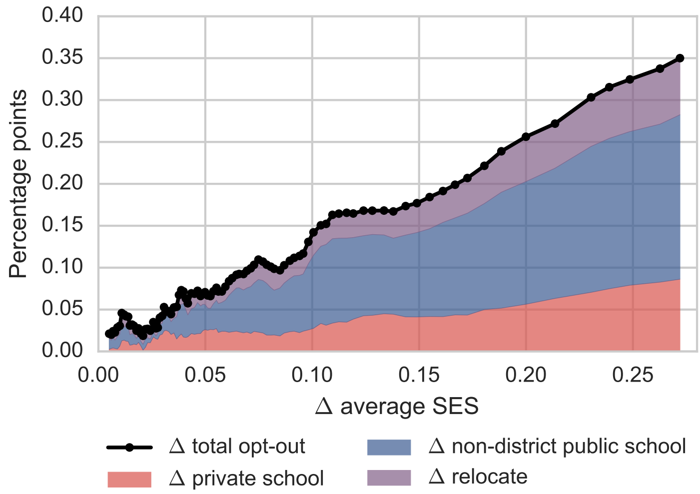


Figure C.3: Differences in opt-out rate as a function of difference in socioeconomic status

The figure depicts the estimated border differences in opting-out of the district school as a function of the border differences in average SES for adjacent district schools. The markers are the γ^+ of the model presented in (7) where we decompose the opting-out into its two subcategories, other public school or private school. The model is estimated in sliding ten percentage point windows of the ranked border difference distribution. Each marker represents the middle of the sampling interval.

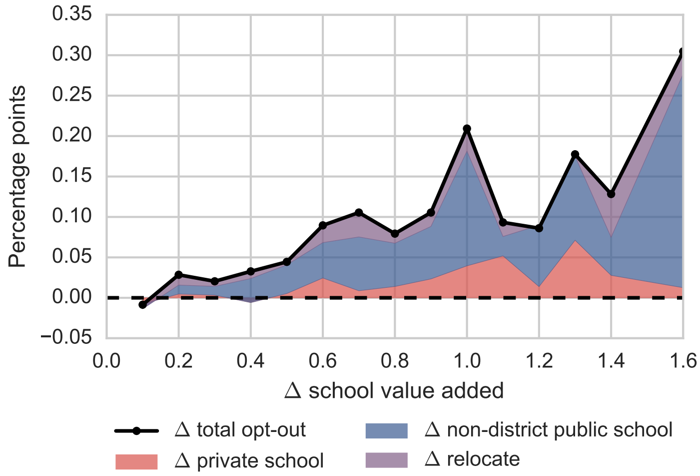


Figure C.4: Differences in opt-out rate as a function of difference in School Value Added

The figure depicts the estimated border differences in opting-out of the district school as a function of the border differences in School Value Added (SVA) for adjacent district schools. The markers are the γ^+ of the model presented in (7) where we decompose the opting-out into its two subcategories, other public school or private school. The model is estimated in sliding ten percentage point windows of the ranked border difference distribution. Each marker represents the middle of the sampling interval.

	Any	Q1	Q2	Q3	Q4
Opt-out of district	-0.118^{***}	-0.069^{***}	-0.131^{***}	-0.116^{***}	-0.131^{***}
	(0.016)	(0.020)	(0.019)	(0.021)	(0.019)
Non-district public school	-0.070^{***}	-0.059^{**}	-0.063^{***}	-0.064^{***}	-0.059^{***}
	(0.014)	(0.018)	(0.013)	(0.014)	(0.012)
Private school	-0.025^{***}	-0.004	-0.038^{***}	-0.028^{*}	-0.051^{***}
	(0.007)	(0.008)	(0.009)	(0.011)	(0.015)
Relocation to new district	-0.024^{***}	0.002	-0.029^{**}	-0.026^{**}	-0.025^{*}
	(0.006)	(0.011)	(0.010)	(0.009)	(0.010)

†: $p < .1$, ∗: $p < 0.05$, ∗∗: $p < 0.01$, ∗ ∗ ∗: $p < 0.001$

Table C3: IV estimates of difference in opt-out rate from difference in School Value Added

The table presents estimations of the marginal propensity to opt out as a function of difference in School Value Added (SVA). School SVA is instrumented by a dummy for being at the “high side” of the district border. Models are estimated with only one quality measure at a time. Standard errors are clustered by the connected component of a graph where an edge exists between border regions if the same household forms part both regions. This procedure creates in 480 clusters.

Chapter 3

The price of free schools

The price of free schools

Andreas Bjerre-Nielsen* & Mikkel Høst Gandil†

Abstract

A central feature of Scandinavian welfare states is the provision of equal access to free primary education. However, if school performance is reflected in property values, economic inequality may diminish equal access. Using highly detailed geographical data for the universe of sales in Denmark in a boundary discontinuity design, we show that property values reflect the socioeconomic composition of student bodies of primary schools in Denmark. Because attendance boundaries change over time we can validate our baseline estimates under less restrictive assumptions and inspect adjustment dynamics. We document that prices begin to adjust immediately and are fully converged within three years. The estimates indicate that our baseline estimates are not inflated by omitted variable bias. Lastly, we calculate that low income households have to forego between 7 and 10 percent of consumption in order to gain access to a socioeconomically strong school. Our findings underline that even when primary school funding is centralized, there are severe obstacles to ensuring equal access to education regardless of parental background.

1 Introduction

Public provision of primary education is one of the main tasks of local government. While Tiebout (1956) showed that local government may

*University of Copenhagen, andreas.bjerre-nielsen@econ.ku.dk

†University of Copenhagen and the Economic Council of the Labour Movement, mga@econ.ku.dk

allow for optimal provision of a public and localized good such as education, a theoretical literature has shown that externalities in educational production can lead to inequality and economic inefficiency (Benabou, 1993; Durlauf, 1996). A crucial mechanism is property prices. If prices reflect educational quality, low-income households may be unable to buy access to good schools, which in turn create inequality of opportunity.

In this paper, we provide evidence on the effect of school characteristics on real estate prices in Denmark, a welfare state with extensive transfers between local governments. Contrary to the American system, the transfers imply that variation in local house prices is not determining school funding to the same degree. Differences in school performance may, therefore, operate more through externalities in education than through funding.

We begin by exploiting cross-sectional variation in school characteristics around attendance boundaries, an approach we refer to as boundary discontinuity design. To decrease omitted variable bias from household sorting around boundaries, we make use of extremely detailed individual geographic administrative data to control for the socioeconomic composition of the 200 nearest adults to a given house sale. Once we introduce these controls, we find that the magnitude of the estimates falls considerably. We show that a standard deviation increase in the school average socioeconomic index is associated with a price increase of between 1.4 and 6 percent, with 3.6 as our preferred estimate. The neighborhood characteristics both affect and are affected by school characteristics. The control variables are therefore partly endogenous and therefore bad controls. We, therefore, argue that the low estimates represent a lower bound of the true effect and that the severity of bad controls is under-appreciated in the literature at large.

We investigate the identification strategy underlying the boundary discontinuity design by recasting it as an IV estimator with a single instrument, a binary variable for a sale being on a specific side of a boundary. Because there is only one instrument, we argue that the exclusion restrictions necessary for interpreting partial estimates of multiple measures of schools in the same regression are unduly restrictive. Such estimates might thus be of little use. Empirically, we show this by replacing the socioeconomic index with ethnic composition as a measure of schools. Due to the very high correlation between the two school measures we get almost indistinguishable results. We argue that the two variables may measure the same underlying latent socioeconomic index. Furthermore, if households do not make the distinction between the two measures, it makes little sense for the econometrician to try to estimate partial effects.

Using school GPA averages we find evidence of positive effects on prices. However, this result is neither robust to the inclusion of neighborhood controls nor controls for the socioeconomic composition of schools. This mixed evidence is somewhat inconsistent with the international literature at large. We show that time variation in GPA within schools is much larger than for socioeconomic variables and therefore argue that parents may use socioeconomic factors as a better proxy for performance.[1]

The boundary discontinuity design is based on cross-sectional variation across geographic space. We complement the existing literature by exploiting time variation in the shape of attendance boundaries. These changes provide exogenous variation in school characteristics, which is not subject to sorting bias. In a difference-in-difference framework, we compare households, which end up in the same district after a change. The estimates from this approach are in line with the cross-sectional estimates and somewhat higher than our estimates using the neighborhood controls. This supports our conjecture that neighborhood variables cause attenuation bias due to being partly endogenous. We are therefore confident that we identify a lower bound of the effects of school characteristics on house prices.

With the changes in boundaries, we investigate the time profile of adjustment. We find evidence that prices begin to adjust within the same calendar year after the reallocation to a new school. After three years the prices are at the same level as for houses which have been assigned to the same school throughout.

Our results are important for understanding the implications of residential sorting on equality of opportunity as richer households may buy

[1]This is consistent with Kane et al. (2003) who argues that yearly school test scores have poor predictive power on house prices.

better educational inputs to the detriment of poorer households. By back-of-the-envelope calculations, we find that a household at the tenth percentile in the income distribution must give up between seven and ten percent of disposable income to move from the tenth percentile to the ninetieth percentile of school socioeconomic composition. This is a major consumption loss and therefore likely to create inequality in educational inputs for children, depending on parental resources.

Attendance boundaries may make the importance of geography larger than need be. The literature on matching and school choice has developed other allocation mechanisms which allow for admittance criteria to be less geographically dependent. However, schools are local in nature as children can only travel a limited distance to receive primary education. We, therefore, argue that geography is a binding constraint and that allocation mechanisms may have to be accompanied by housing and zoning policies to ensure equal access to education.

We begin by reviewing the literature and the theoretical reasons why house prices may reflect valuations of school characteristics in section 2. We then proceed to describe the institutional context of allocating student to Danish primary schools in section 3 and describe our data in section 4. The analysis is presented in section 5 and we perform a back-of-the-envelope calculation of the distributional implications of our results in section 6. Section 7 concludes.

2 Methods and literature

The theoretical link between schools and house prices has been studied in detail by Benabou (1993); Epple and Romano (1998); Durlauf (1996) among many others. These studies investigate the link between residential sorting and the financing and provision of educational services as a local (club) good. An important theoretical link goes through externalities in the educational production function. A simple argument goes as follows: If children benefit from exposure to other children with high human capital and if families with high human capital tend to have more financial resources, then rents will reflect these positive externalities. As high-income families are able to pay more for housing, the low-income families will not be able to buy access to these communities and thus are excluded from the beneficial exposure to strong peers.[2] Importantly, this dynamic can exist even if funding is centralized. The link between schools and housing prices is therefore crucial for the possibility of securing equal access to education.

Empirically, the dominant approach to estimate the importance of schools on prices is to conceptualize a home as a composite good, following Rosen (1974). In equilibrium, such a model equates the price of an amenity to the valuation of the marginal buyer. This is the theoretical underpinning of an attempt to model a hedonic price function, which links local amenities to marginal valuations via house prices.[3]

As with almost all empirical economics, the issue of simultaneity and omitted variable bias is a great concern in this literature. The close links between schools, prices, sorting and local political economy create reverse causality, so that house prices exclude low-income groups from buying houses in a school attendance zone, thereby affecting the student body of the local school and possibly the level of funding. As noted by Black and Machin (2011), empirical research is limited in means to control for this type of effects.[4]

Local amenities cause bias insofar as they correlate with school characteristics and are not included in the regression. These amenities could be physical facilities, such as sports and recreational facilities and access to public transport. Early studies mostly sought to expand the number of controls to reduce the bias, see Kain and Quigley (1970) for an example of this approach. However, the general literature has moved towards trying to reduce omitted variable bias using a more reduced form approach where the source of identifying variation is made explicit. Chief among

[2]See Benabou (1994) for a more rigorous version of this argument.

[3]While the Rosen model most directly lead to a structural econometric approach, Black and Machin (2011) note, that the model forms the basic underpinning of most of the literature on the housing market and schools, though sometimes implicitly.

[4]As argued by Black and Machin (2011) this is especially a problem in the U.S. context, where schools are funded via property taxes. In our context, we investigate within municipality variation, whereby funding is not a problem. However, a nuance is that we in essence measure the after-tax valuation of school characteristics.

these approaches is the Boundary Discontinuity Design (BDD), which we also employ in this paper. The identifying assumption for this approach to yield unbiased estimates is that unobserved amenities vary continuously while school characteristics are determined by attendance zones, and thus are discontinuous at boundaries. The unobserved amenities thereby cancel out as they are shared across borders.

Black (1999) was the first to use this approach for school districts in Massachusetts. The number of papers later using this approach are legion. Estimates from BDD are typically five times lower than cross-sectional estimates which shows that unobserved heterogeneity is important in hedonic pricing models, as documented by Kane et al. (2006); Bayer et al. (2007); Gibbons et al. (2013) among others. A central issue is that school characteristics are not the only thing that changes discontinuously at the border. Bayer et al. (2007) show that estimates in a BDD-framework change substantially once additional controls are included, implying that neighbors and house characteristics also tend to vary at the border. Gibbons et al. (2013) approach this problem with a myriad of control techniques, such as weighting and spatial trends.

A range of papers has shown the importance of other factors for the valuation of schools in a BDD framework. Fack and Grenet (2010) match sales across borders and show that access to private schools diminishes the importance of public school characteristics in Paris. A number of studies explore whether the release of public information about test scores and school-value-added affect the capitalization of school quality in house prices, see Kane et al. (2003); Figlio and Lucas (2004); Kane et al. (2006); Fiva and Kirkebøen (2011); Imberman and Lovenheim (2016).

While the BDD approach has been very popular, the issues with sorting around borders have led to other strategies exploiting temporal variation in either school characteristics or assignment to identify the valuation of schools. Changes to school boundaries have been explored with difference-in-difference approach investigating one-time changes to local school attendance boundaries in Shaker Height, Ohio., U.S. (Bogart and Cromwell, 2000) and Vancouver, Canada (Ries and Somerville, 2010). A drawback in both of these studies is that they have no information about neighborhood quality and school composition and use a single redrawing of the boundaries.[5]

Most studies use test scores as a measure of schools. Quantitative results from numerous studies are reported by Black and Machin (2011). The authors conclude that the baseline estimate is that a standard deviation increase in test scores increase house prices by 3 percent.

2.1 The Boundary Discontinuity Design

We now present the Boundary Discontinuity framework in more detail. We begin by presenting a simple model and then recast the estimator as an Instrumental Variable (IV) approach to elucidate the necessary assumptions to provide causal estimates.

We assume that log house prices of dwelling i is a function of a school characteristic, q_s. Suppressing the time-dimension we write:

$$p_{is} = \kappa + \beta q_s + u_{is}, \tag{1}$$

where κ is a constant. Under the assumption that $E[q_s u_{is}] = 0$ we can estimate (1) by regressing p_{is} on q_s. However, the moment restriction can be violated for all sorts of reasons. To the extent that the composition of the housing stock and local amenities correlate with the school characteristics, a simple regression according to (1) will yield biased results.

A first approach is to amend (1) with additional controls for housing characteristics and measures of local amenities. However, in order to yield an unbiased estimate of β, these controls must be exhaustive. A central worry is the role of unobserved amenities, which may correlate with school characteristics. However, as Black (1999) noted, if schools vary discontinuously while unobserved amenities do not, then by comparing houses close to each other, but on either side of the border, unobserved amenities cancel out. In other words, if children are allocated to schools via attendance boundaries (SAB), then this creates a discontinuous jump in school characteristics at the border of two SABs. This insight leads to

[5]Ries and Somerville (2010) address this problem using a repeated sales price kernel Additionally, Bogart and Cromwell (2000) acknowledge that their sample is small and limited to only high-quality schools.

the Boundary Discontinuity Design (BDD).[6] To put this in formal terms, we add a boundary fixed effect to (1) and exclude the constant term:

$$p_{isb} = \beta q_s + \mu_b + u_{isb}, \tag{2}$$

With the fixed effect we exploit only variation within a border region, b. This implies that we control for all characteristics shared among houses on *both* sides of the border, whether they are observed or not. If $E[q_s u_{isb}|\mu_b] = 0$ we can estimate (2) by way of OLS. To discuss the validity of such assumptions, it is useful to reframe (2) as an IV estimator. Define a dummy r_i, which takes the value of 1 if house i is on the high side of the border b. Under the same assumptions leading to unbiased estimates of (2) it must hold that $E[r_i u_{isb}] = 0$. We can therefore calculate the Wald estimator as:

$$\beta^{Wald} = \frac{E[p_i|r_i = 1] - E[p_i|r_i = 0]}{E[q_i|r_i = 1] - E[q_i|r_i = 0]} \tag{3}$$

Observe that the reduced form, i.e. the nominator in (3), is the average difference in prices across borders. For this to be a valid estimate, we need r_i to uncorrelated with other variables, such as (unobserved) neighborhood characteristics. That is, we cannot allow sorting across the border. In other words, the exact position of the border should be as good as random. This is the standard Regression Discontinuity assumption, that the distribution of covariates is continuous at the discontinuity, see Imbens and Lemieux (2008). If this assumption holds, then the nominator of (3) is the average effect of being on the "high" side of a district border. However, this is not a very useful measure in and of itself, as it needs to be rescaled by the first stage to provide an estimate of a marginal effect. The first stage is the denominator in (3). For the Wald-estimator to yield an unbiased estimate, the exclusion restriction needs to be valid. In other words, r_i must only affect p_i through its effect on q_s. The formulation of the model as a Wald estimator highlights the severity of the exclusion restriction in this framework. In (2) schools are measured by a scalar,

[6]The Boundary Discontinuity Design is equivalent to a Regression Discontinuity Design with distance to a border as the running variable.

q_s. As schools exhibit multiple characteristics, we should ideally have an instrument for each characteristic. Nevertheless, we only have one: the dummy for crossing the border. If being on the high side according to one school measure correlates with being on the high side on some other school measure, then omitted variable bias is still an issue.

Though the inclusion of the border fixed effect removes unobserved factors shared at the border, it does not ensure identification of causal *partial* effects. Thus, the effect of different socioeconomic factors may not be causal when included jointly, as is normally done in the literature. In our data, we observe very strong correlations between a socioeconomic index and ethnicity.[7] Based on the outline above, however, we do not believe that the two factors can be separately identified because both may proxy for the same underlying but unobserved socioeconomic factor. We can therefore use either measure as a proxy for this underlying factor, but not both.[8] Without knowing the underlying correlations between measures, this leaves us little confidence in "horse race" type regressions where multiple measures are included to see which factors explain the most.[9]

Bad controls? As mentioned above, a fundamental issue in BDD is sorting across the boundaries. This is especially a problem with school attendance boundaries. It might be that the marginal buyer considers neighbors with high socioeconomic status (SES) a valuable amenity. Thus, failing to include a measure of the neighbor-composition will bias the estimate of schools. Including the neighbor-composition may, however, cause problems as well. If high-SES schools attract high-SES households,

[7]See section 4 and appendix A

[8]To see why we do not identify the partial effects, observe that children both have an ethnicity and a socioeconomic index. In other words, separate variation in these two characteristics requires different compositions of children. However, a socioeconomic group with a high minority-share is fundamentally different from another group with the same socioeconomic composition but a different minority share. Thus, any unobserved differences between the two groups will bias the partial results.

[9]While we doubt the feasibility of estimating regressions with more than one school characteristic at a time, we do provide results in appendix C in order to ensure comparability with results from the hedonic literature. In the main text, we will not interpret these joint measures.

then the neighborhood-SES is a function of the school and will therefore be a "bad control". However, high-SES households will also send their kids to the local school thereby increasing the school SES further. The measures of schools and neighborhoods are therefore completely intertwined. Nevertheless, we may be able to bound the effect in a simple model.[10]

Suppose that a_i is an amenity of house i (or its' vicinity). We do not observe a_i but have a proxy, $\tilde{a}_i$, which is some function of the school characteristic and the true amenity; $\tilde{a}_i = \pi_0 + \pi_1 q_s + \pi_2 a_i$. Now assume that the true model, instead of (2) is given by the following, where we exclude the fixed effects for convenience:

$$p_{isb} = \beta q_s + \delta a_i + u_{isb}. \tag{4}$$

We assume that $\delta > 0$, $\pi_1 > 0$ and $\pi_2 > 0$. Without including the local neighborhood variable, we would estimate $\hat{\beta}^+ = \beta + \delta\lambda$, where $\lambda = Cov(q_s, a_i)/Var(q_i)$. This estimate is biased upwards if the amenity correlates positively with the school characteristic. The alternative is to include the neighborhood variable:

$$\begin{aligned} p_{isb} &= \kappa + \beta q_s + \delta q_i + u_{isb} \\ &= \kappa + \beta q_s + \delta \left(\frac{1}{\pi_2} q_i - \frac{\pi_0}{\pi_2} - \frac{\pi_1}{\pi_2} q_s \right) + u_{isb} \\ &= \left(\kappa - \delta \frac{\pi_0}{\pi_2} \right) + \left(\beta - \delta \frac{\pi_1}{\pi_2} \right) q_s + \frac{\delta}{\pi_2} \tilde{a}_i + u_{isb}. \end{aligned} \tag{5}$$

If we include the neighborhood variable, we will therefore estimate $\hat{\beta}^- = \beta - \delta \frac{\pi_1}{\pi_2}$, which is negatively biased under the parametric assumptions. In other words, in this simple example we can bound the true effect, β, by estimating models with and without hyper-local neighborhood controls.

If we run a regression of (5), we see that the parameter on the proxy for the amenity is $\frac{\delta}{\pi_2}$. In other words, if we do not observe the true amenity but only the proxy, we cannot identify δ unless $\pi_2 = 1$. Our attempt at bounding β should lead us to respect the fundamental un-

[10] In this example we follow Angrist and Pischke (2008) closely.

certainty about the relation between the proxy amenity and the "true" amenity. If we were to interpret on the magnitude of the coefficient on our proxy, we assume knowledge of the true underlying parameter. We caution that the bad control problem causes issues with both the parameters on schools and the proxy for the unobserved amenity. Due to these considerations, we take a less structural approach to interpretation of parameters on neighborhood characteristics than what is mostly done in the hedonic pricing literature.[11]

2.2 Adjustment to changes in boundaries

Attendance boundaries change from time to time. This entails a shock to the school characteristics for some addresses. In the short term, however, other amenities should be approximately constant. If the estimates we find using the Boundary Discontinuity Design are causal, one should expect effects of the SAB change on the house prices to be of the same sign and magnitude. To investigate this, we employ a difference-in-difference approach. We focus on the closing of a price gap and we therefore reverse the time dimension compared to the usual difference-in-difference approach. For school s, we look at dwellings transferred *into* the corresponding SAB. Let r_i be an indicator for whether the address is transferred from one SAB to another at some point and let τ_i be the year the dwelling is transferred. We include a fixed effect for all "arrival SAB"-year combinations and run variations on the following regression:

$$p_{ist} = \theta \cdot r_i \times \mathbf{1}(t < \tau_i) + \lambda r_i + \mu_{st} + u_{its} \tag{6}$$

The fixed effect ensures that we are only using differences within a year within the attendance boundary where all dwellings end up. The parameter λ picks up the time constant difference between those dwellings that are transferred and those who are not. The parameter on the interaction, θ, measures the change in differences before and after the change of the

[11] As an example, Bayer et al. (2007) interpret the change in parameters on neighborhood characteristic before and after the inclusion of school characteristics as a valuation of neighbors over and above what their effects on school characteristics.

attendance boundary. Naturally, a transfer to a new school can entail a fall or an increase in school characteristics depending on the departure schools. To accommodate this, we let r_i take the value of negative one if a dwelling is transferred into a lower measured school.[12]

The specification in (6) measures the change when transferred from a "low" to a "high" school. This is equivalent to the reduced form found in the BDD-framework. To get at the marginal effects we can reformulate (6) by including the school characteristic as a continuous variable:

$$p_{its} = \psi q_{s't} + \lambda r_i + \mu_{st} + u_{its}, \tag{7}$$

where $q_{s't}$ is the school characteristic in year t of the school s' in the same year. Once again, λ picks up time-constant differences between those transferred and those not transferred, whether they be unobserved or not. After the change in boundaries, school characteristics are the same for all dwellings within the SAB of school s - whether transferred into it or not. Because of the fixed effect, the only variation in school characteristics, therefore, comes from changes in SAB across borders *before* the boundary change and ψ provides an estimate of the effect of school characteristics on prices which we can compare to our BDD estimates. Prices may not adjust instantly, and we, therefore, investigate the timing of responses in the analysis.

The specifications above are simplified for exposition. As we estimate the regressions, we elaborate further on how we control for observed covariates and neighborhood characteristics.

3 Primary schools and attendance boundaries in Denmark

Danish primary schools are run by municipalities who decide how to prioritize the general level of funding according to their full set of priori-

[12] This is equivalent to estimating the model for positive and negative shock and then adding up the "flipped" results, such that the treatment from a negative shock is given a negative value.

ties.[13] Schools are free and parent co-payment is forbidden by law. This implies that school funding generally does not vary within municipalities. Students are allocated to schools via residential zones, referred to as school attendance boundaries (SAB), districts or catchment areas.[14] Municipalities can change these boundaries as they wish. Anecdotally, they do this due to projected capacity constraint and development of socioeconomic compositions of schools. Administrative authorities in the municipality usually announce changes in boundaries within a year of implementation. If a child lives within a given boundary, she is guaranteed enrollment in the associated school. If a school is not fully subscribed it is possible for children from other districts to be enrolled. Thus, the living within an attendance boundary is a guarantee, but not a determinant for enrollment into a given school. Bjerre-Nielsen and Gandil (2018a) describe the rules in further detail. Municipalities are financed by income taxes and land taxes which are in general set within a tight bound.[15] An extensive system of transfers between municipalities ensure the municipalities with high expenses due to sociodemographic factors are compensated by other municipalities.

4 Data and descriptive statistics

In this section, we briefly review our data and provide descriptive statistics. Much of this section reflects the same choices and restriction used in Bjerre-Nielsen and Gandil (2018a).

House sales Our main dataset comprises sold dwellings from 2008 to 2015 as provided in the dataset EJSA by Statistics Denmark. We observe price, a date of sale and an identifier, EJENDOMSNUMMER. We link this identifier to individual owners through the dataset EJER. We then link the

[13]Among other things, Danish municipalities are also responsible for child and elderly care, environmental protection, urban planning and execution of active labor market policies.

[14]In Denmark school districts contain only one school. We, therefore, use the three terms interchangeably.

[15] In 2015 the municipal taxes ranged from 22.5 to 27.8 percent of income, with 50 percent of municipalities within 24.9 and 25.8 percent.

sales to addresses by merging this data to administrative records on individuals living in their own home from 1990 to 2016. To obtain detailed geographical data we link the addresses to a modified version of the Danish Squarenet. This dataset consists of very small polygons (100×100 meters in densely populated areas). We define the location of a dwelling as the centroid of the associated polygon. This provides detailed geographic location while maintaining a degree of anonymity, see Bjerre-Nielsen and Gandil (2018c) for documentation. We remove addresses for which we do not have a sufficient degree of precision.[16] We exclude farmhouses and sales where building type is not observed.

School attendance boundaries We obtain school districts for each year from the Central Person Registry (CPR) which publishes a file every quarter with school districts provided by the municipalities.[17] We have data dating back to the nineties. However, due to a municipal reform in 2007, we limit our sample to the period 2008-2015. The districts are voluntarily provided by the municipalities and are not subject to any quality checks. We remove a few municipalities where there are obvious errors in the reporting.[18] The districts are provided as lists of addresses. We clean this data and merge the addresses onto a GIS-dataset containing the spatial features of all plots with associated addresses in Denmark (Martrikelkortet) provided by the Danish Geodata Agency. We define the attendance boundaries as the edge of the spacial polygon made out of the union of the plots in the district. We calculate distances as the shortest Euklidean distance to the boundary from the centroid of the residential polygon. In the boundary discontinuity analysis, we include only addresses with a distance of less than 2000 meters to the borders.[19] To avoid unobserved differences in taxes and provision of public services we only include dis-

[16]Specifically we remove polygons where the share of the area to the convex hull is below 0.4.

[17]As the CPR overwrites previous versions we a grateful for the due diligence of a retired employee in Statistics Denmark, as this has been the only backup known to us.

[18]These municipalities include Vordingborg and Bornholm. We also exclude Gentofte, as it uses fluid attendance boundaries, which likely is internalized into the location decision of households and thus prices.

[19]We restrict the distance restriction further in parts of the analysis.

tances to borders within the same municipality.

School characteristics We apply three different school characteristics. The first characteristic is the mean household socioeconomic index (SES) which we define as the first component a Principal Component Analysis using household income rank and dummies for education and employment of the adult members of the households. We rank the component such that the index is uniformly distributed. The index aligns closely with our intuition that high-SES households are employed, educated and have high incomes. For further detail, we refer to the appendix of Bjerre-Nielsen and Gandil (2018a).

For each school in each year, we calculate the average household-SES of the enrolled students. We call this average the school-SES. In the analysis, this will be our main measure of schools. We also calculate the share of Non-Western immigrants and descendants enrolled in the school. We include this measure as it is very prominent in the public debate in Denmark.[20] The last characteristic is the graduating average of ninth graders (I.e. the final year of public school). This average is public information and thus freely available to prospective house buyers from a website. We standardize the school grade averages, such that they have mean zero and a unit standard deviation.[21]

Neighborhood characteristics We link individuals to the Squarenet in order to calculate hyper-local neighborhoods; For every polygon, we calculate the mean of socioeconomic characteristics for the 200 nearest adults in each year. These characteristics are the household-SES index and dummies for employment, long-cycle education and non-Western immigrant or descendant.

[20]As we explain below, we observe such a strong correlation in our data between having a high non-Western share and a low school-SES that we do not feel confident in disentangling the two characteristics.

[21]We calculate the z-score of GPA for all available data. However, our sample restrictions cause the mean to be positive and the spread to be smaller than one as evidenced in table 1b.

Descriptive statistics Descriptive statistics for our sample of sales are presented in Table 1a. The first column display results for the total sample. The following two columns split the sample into two, whether the dwelling is on the low or high side of a border, measured by school-SES. It is evident that the sample is not balanced; on the high side, there is a larger share of single-family homes and higher mean square footage. Conversely, houses on the low side are much more likely to be apartments. These differences are not surprising, insofar as high-SES households are more likely to reside in larger houses and at the same time send their children to the local school. In addition, the neighborhood variables show some unbalances.[22]

In the two right-most column, we calculate statistics for those dwellings, which experience a change in school associations sooner or later and the control group, which maintain the same school association throughout. The transferred dwellings seem overall comparable to the control group. However, there is a larger share of apartments and a higher level of education among neighbors. Half of the sales of eventually changed properties occur prior to the change, reflecting that most of the boundary changes occur around 2011. In Appendix Figure A.1 we report cumulative distributions functions for distances to borders.

We present school level descriptive statistics in Table 1b. As we employ schools in multiple years, we also present the share of variance within school relative to the total variance. This gives us a sense of the stability of school characteristics over time. For school SES and non-Western this share is very low, while almost 40 percent of the variance in GPA stems from variation within schools. Insofar as the socioeconomic composition of schools affects or correlate with student performance, families may regard these socioeconomic statistics as better indicators of school quality than the GPA itself. We do not have data on all schools, and the sample size, therefore, drops whenever GPA is used.

In Appendix Figure A.2 we plot the joint distribution of border dif-

[22]As previously mentioned these neighborhood variables might be thought of as bad controls, and we will discuss the implications of including them as controls in the analysis.

(a) Dwelling level

Variable	Statistic	All BDD	Low side SES	High side SES	SAB changed	SAB control
M^2	mean	123.04	118.57	127.18	124.76	126.98
	median	119	114	124	120	124
	std	43.05	42.06	43.54	45.90	43.31
Distance to border	mean	880.52	874.21	886.38		
	median	823	814.50	831		
	std	534.07	536.44	531.79		
N: Employment, share	mean	0.79	0.78	0.80	0.80	0.79
	median	0.80	0.79	0.81	0.81	0.81
	std	0.08	0.08	0.08	0.08	0.08
N: Long-cycle education, share	mean	0.11	0.10	0.12	0.16	0.10
	median	0.08	0.07	0.09	0.13	0.07
	std	0.10	0.09	0.10	0.12	0.10
N: Non-Western, share	mean	0.05	0.06	0.04	0.06	0.05
	median	0.03	0.04	0.03	0.04	0.03
	std	0.06	0.07	0.06	0.06	0.06
N: SES, mean	mean	0.55	0.54	0.57	0.58	0.55
	median	0.55	0.54	0.57	0.59	0.55
	std	0.08	0.08	0.08	0.09	0.08
Single-family home	share	0.60	0.56	0.64	0.54	0.66
Terraced house		0.18	0.18	0.17	0.17	0.16
Apartment		0.22	0.26	0.18	0.29	0.18
Before change	share				0.49	
Obs.	Count	341697	164344	177353	11161	383928

(b) School level

	Mean	Median	Std.	N, school-years	N, schools	Within/Total variance
School SES	0.48	0.47	0.11	7473	1357	0.03
Non-Western share	0.11	0.06	0.13	7473	1357	0.02
GPA	0.20	0.21	0.67	5714	984	0.38

Table 1: Descriptive statistics

The table presents selected statistics from the total BDD sample. The prefix "N" denotes that it is a hyper-local neighborhood variable. Only observation under 2000 meters from the border is kept four the boundary discontinuity analysis. We do not impose this restriction when investigating changes in SAB. "Before change" is the share of sales of properties, which are eventually transferred to another SAB but which are observed prior to the transfer. In the analysis, the sample is restricted in different ways described in the text.

ferences in school SES and the non-Western share. The correlation is close to -0.9. As we repeatedly state in this paper, we believe that we are not able to identify the effect of ethnicity and socioeconomic status separately with a correlation of this magnitude. We thus doubt the validity of insights from partial effect estimates of SES and non-Western share, holding the other constant. We return to this point in the next section.

5 Analysis

This section is split in two. We first present evidence using a boundary discontinuity design in section 5.1 and then proceed to present the results from the changes in attendance boundaries in section 5.2.

5.1 Static border comparison

We begin by estimating the discontinuities at school borders for the three measures of school, one at a time. First, we construct bins 200-meter bins of distance to the boundary. As is conventional in the literature, we define distances from a border as negative if the address belongs to the school, which has the lower measure of the two schools sharing the border. We then run regressions of the following kind:

$$p_{ibt} = \sum_{d=d^-}^{d^+} \lambda^d \mathbf{1}(dist = d) + \mu_{bt} + \mathbf{X}_{it}\eta + \mathbf{Z}_{it}\delta + \varepsilon_{isbt}, \tag{8}$$

where p_{ibt} is the log price of house i at the border b sold in year t. The border-year fixed effect, μ_{bt} insures that we take out any level differences, shared by the two sides of the border.[23] We include a vector of dwelling characteristics, size and building type, in $\mathbf{X}_{it}$. We also include polynomials of the hyper-local neighborhood variables in the vector $\mathbf{Z}_{it}$. As we do not condition the hyper-local neighborhoods to be on either side of the boundary, we implicitly control for spatial trends across the border by in-

[23]Note, that sales may be in multiple border regions and thus enter as multiple observations which introduce issues of serial correlation. We therefore cluster standard errors at municipal level.

cluding $\mathbf{Z}_{it}$. We only estimate (8) for border regions where the difference in school characteristics exceed a standard deviation of the total border difference distribution.[24]

In Figure 1 we plot the estimated λ^ds from the model in equation (8) with SES as the measure of schools. The black points are from an estimation without controls and show a clear discontinuity, with a mean difference of 0.16 points. However, much of this price difference is due to differences in house characteristics. When we include housing characteristics, the difference shrinks to seven points. When we include the hyper-local neighborhoods and house characteristics jointly, the discontinuity falls to three points. This is substantially lower than without controls, though still highly significant.

We perform the same exercise for non-Western share and the official GPA-average and exclude the raw estimates without controls as they dwarf the other estimates in magnitude and hinder visual inspection of the discontinuities. Figure 2a repeats the estimation from Figure 1. In Figure 2b we find mimicking results when we measure schools by their non-Western share instead of average SES, though the trends are noisier. The average difference is four log points which decreases to two points when we introduce neighborhood controls. Again, we stress that school-SES and the non-Western share are highly correlated. Lastly, we measure schools by the average GPA and display the results in Figure 2c. We find a discontinuity of five log points in the estimation with house controls. This result falls to two log points, once we introduce neighborhood controls and the visible discontinuity disappears. As mentioned in section 2, the hyper-local neighborhood may be a bad control insofar as the local sociodemographic makeup is a function of school characteristics. We, therefore, regard these estimates as a lower bound on a true effect. Nevertheless, even if we shut down the effect from neighbors we find significant, positive effects of socioeconomic stronger schools on house prices.

[24]We only make this restriction for plotting the discontinuities in Figures 1 and 2. In Appendix Figure D.1 we investigate whether the difference in school-SES matter for estimating marginal effects.

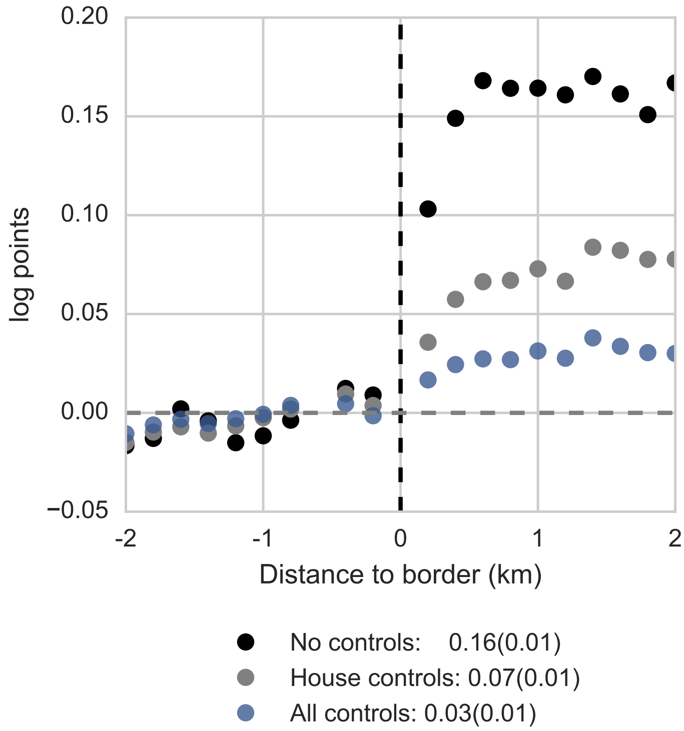
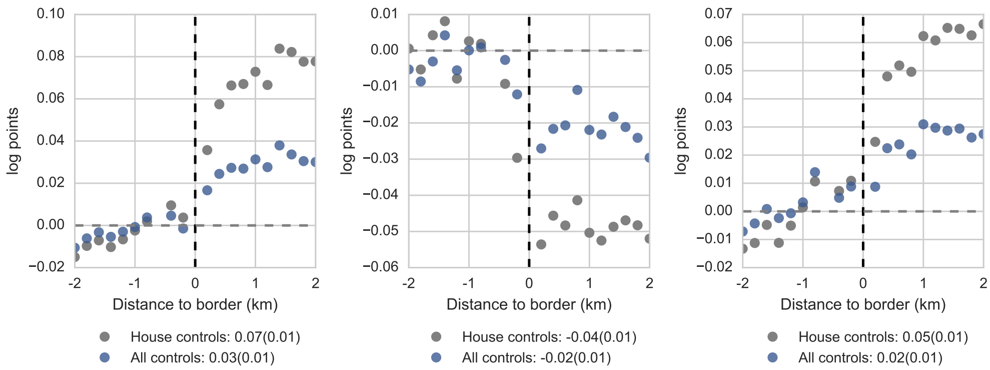


Figure 1: Simple Boundary Discontinuity Design

The graphs show the results of a Boundary Discontinuity Design with SES as school measure. We run regressions of discretized distances to the border, where negative distances signify the address belong to the side of the border with the lowest value of the measure in question. We include a border-year fixed effect to control for level differences shared by both sides of the border. In black, we present the parameters on the binned distance dummies with no additional controls beside the fixed effect. The parameters in gray are from an estimation where we include hyper-local neighborhoods and square meters (including all controls squared). The results are normalized at the 400-meter distance bin at the left side of the border. We also compute an average difference by regressing log prices on a dummy for being on the right side and border-year fixed effects for the same sample. We display the parameters from these regressions in the legend along with standard errors, clustered at the municipal level, in parenthesis. We only include borders with a difference in school SES over one standard deviation. In Appendix Figure B.1 we display corresponding figures with confidence intervals on the binned distance dummies.



(a) School SES (b) Non-Western share (c) Average GPA

Figure 2: Simple Boundary Discontinuity Design

The graphs show the results of a Boundary Discontinuity Design with one school characteristic at a time. We run regressions of discretized distances to the border, where negative distances signify the address belong to the side of the border with the lowest value of the measure in question. We include a border-year fixed effect to control for level differences shared by both sides of the border. In black, we present the parameters on the binned distance dummies with no additional controls beside the fixed effect. The parameters in gray are from an estimation where we include hyper-local neighborhoods and square meters (including all controls squared). The results are normalized at the 400-meter distance bin at the left side of the border. We only include border regions where the difference is above a tenth of standard deviation of the "border difference distribution". We also compute an average difference by regressing log prices on a dummy for being on the right side and border-year fixed effects for the same sample. We present parameters from these regressions in the legend along with standard errors, clustered at the municipal level, in parenthesis. In Appendix Figure B.1 we display corresponding figures with confidence intervals on the binned distance dummies.

5.1.1 Estimating marginal effects

The discontinuities in Figure 2 provide evidence that schools may causally affect house prices. However, in order to compare these estimates to other results in the literature, we need the discontinuities expressed as marginal effects. We do this by implementing regressions of the following type:

$$p_{ivsbt} = \beta q_{st} + \mu_{vbt} + \mathbf{X}_{it}\lambda + \mathbf{Z}_{it}\delta + \varepsilon_{ivsbt}, \tag{9}$$

where p_{isbt} is the log price of house i in the SAB belonging to s at the border b sold in year t. The school characteristic of school s at time t is measured by q_{st}, and β is the parameter of interest. We include a "house type"-border-year fixed effect, μ_{vbt}. In other words, we are only comparing within housing category within border within year. Due to the fixed effect, all variation in q_{st} comes from crossing the boundary. We restrict our data to be within 300 meters of the district border.[25]

We begin by presenting regressions of prices on one school characteristic at a time where we control for house characteristics, neighborhood characteristics and the border-year-type fixed effects.[26] Column 1-3 in Table 2 present the estimates. We see the same pattern as in figure 2 as school-SES and the non-Western share maintain their signs and significance. A standard deviation of school-SES ($\approx$0.1) causes prices to rise by 1,4 percent. Conversely a standard deviation increase of the NW-share (again $\approx$0.1) cause prices to fall by 0.7 percent. GPA also has a positive effect, but a standard deviation increase in GPA ($\approx$ 0.7) only entails a price increase of 0.3 percent and the effect is very small and insignificant. This effect is an order of magnitude smaller than effects documented elsewhere in the literature.

Column 4 in Table 2 shows the results from regression the three school characteristics jointly. The parameter on a given characteristic is therefore conditional on the other characteristics. The parameter on SES barely changes. The parameter on non-Western share switches sign and falls in

[25] In Appendix Figure B.2 we evaluate the importance of the restriction on distance. We see, that while the restriction is important for the raw estimate in column 1 of 2, it hardly matters once we introduce neighborhood controls.

[26] Types include single family homes, terraced housing and apartments.

	(1)	(2)	(3)	(4)
School SES	0.139***			0.246**
	(0.0345)			(0.0740)
Non-Western share		-0.0733**		0.0561
		(0.0253)		(0.0474)
GPA			0.00432	-0.0119
			(0.00592)	(0.00651)
House controls	X	X	X	X
Neighborhood controls	X	X	X	X
N	56374	56374	52872	52872

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Marginal effect estimates

The models present estimates from an OLS regression of house prices on school characteristics and a border-type-year fixed effects. Standard errors, clustered at municipal level, are presented below the estimates. Only houses within 300 meters of the border enter the regression. For full regression output see Appendix Table B1.

magnitude to a no longer significant effect of 0.03. The parameter on GPA remains essentially zero.[27]

As explained in section 2, we do not believe that our source of variation identifies the partial effects of the three measures. Furthermore, there might be a fourth, unobserved, school characteristic explaining the variation. It is therefore unclear whether it is meaningful to try to distinguish between these variables in the first place. With such high correlations, house-buyers may not make the distinction themselves. The three school measures may essentially reflect the same underlying index from the perspective of the buyers. Thus, trying to separate out the partial effects may be meaningless.[28]

In light of the under-identification of school characteristics, we progress

[27] We note that our estimated null effects of school GPA on local house prices from Table 2 is robust to increasing the maximum distances to the boundary, as long as we include school SES and NW, see Table B2 in Appendix B. However, when excluding these school characteristics and increasing the maximum distance to the boundary to 500m the estimate is borderline significant; when including observations within 1000m there is a strongly significant effect. When further excluding neighborhood controls our estimates of GPA on house prices are significant for all maximum boundary distances. In other words, we find a lower bound of essentially zero, but we cannot rule out, that prices might be affected by the average GPA of the local school. The socioeconomic variables, however, seem to carry more weight than the GPA.

[28] Bjerre-Nielsen and Gandil (2018a) discuss this in further detail.

using only the SES as a school characteristic. In appendix table B1 we report the full set of parameter estimates for all school characteristics. In appendix D we investigate heterogeneity in responses as a function of the magnitude of the differences in school characteristics between neighbor schools. We document that differences in prices are almost linearly increasing in differences in school SES, which imply an approximately constant marginal effect of school-SES on prices. For comparison to results in the hedonic literature, particularly Bayer et al. (2007), we construct hedonic regressions in appendix C, but we stress that these regressions are most likely under-identified.

5.1.2 Threats to identification

The large drops in the discontinuities once we introduce controls indicate that the attendance boundaries are not drawn completely at random. We did not expect this to be the case. However, we feel confident that our hyper-local neighborhoods are sufficient to avoid most of the possible omitted variable bias. If houses differ, such that houses on the "high side" of a border are deemed more desirable and thus more expensive, higher income households will also tend to live in them. The hyperlocal neighborhoods, therefore, control for the unobserved features by proxy. For omitted variable bias to still be an issue, the unobserved characteristics must affect the prices in a way that does not influence the socioeconomic composition of buyers. We find it difficult to construct mental models of such sorting. In this context, the local neighborhood controls, therefore, act as proxies for unobserved heterogeneity in house characteristics. As previously mentioned, these finely grained controls may, however, bias the results towards zero, as they are partly determined by the treatment, i.e. the variation in school characteristics. We, therefore, regard the estimates as lower bounds.

Our findings imply that while studies without access to detailed geocoded data on sociodemographic profiles of residents may overstate the importance of schools, schools are important for house prices. We now move to the second identification strategy of this paper, where we exploit time variation in the shape of SABs to validate our results from the boundary discontinuity design.

5.2 Boundary changes and price adjustments

As discussed in Section 2.2 we perform Difference-in-Difference but reverse time. We, therefore, inspect price adjustments to the *new* school association. We begin by regressing sales prices on dummies for time distance to change, covariates and a "arrival-school"-year fixed effect:

$$p_{ist} = \sum_{j=4}^{4} \lambda^j \mathbf{1}(t - \tau_i = j) + \mu_{st} + \mathbf{X}_{it}\eta + \varepsilon_{ist} \tag{10}$$

We are *not* including a dummy for the time constant differences in prices. In other words, if λ^j is zero this means no difference in the price levels between the dwellings already in the arrival district and those arriving. In order to increase power, we recode dummies from one to minus one if the change in school-SES is negative. Figure 3 presents estimates of equation (10). The years before the change in attendance boundaries the coefficients are below zero and significant, while the estimates are close to zero and insignificant after the change. This implies that the price gap between addresses almost close once the addresses become associated with the same school. The adjustment is quick as the coefficient rises towards zero already in the year where the boundary change takes place. Three years after the change the prices have completely converged. The adjustment period is somewhat sensitive to the inclusion of controls but the immediate jump in prices suggests that the capitalization occurs quickly. The coefficients provide supporting evidence that the border differences identified in section 5.1 causally affect house prices.

We sum up the differences by collapsing the model to a pre- and a post-dummy. We estimate the model for the time span between four years prior and four years after the change, nine years in total. We estimate the model in equation (6) amended with controls:

$$p_{ist} = \theta \cdot r_i \times \mathbf{1}(t < \tau_i) + \lambda r_i + \mathbf{X}_{it}\eta + \mu_{st} + u_{ist}, \tag{11}$$

where once again μ_{is} is a fixed effect for arrival SAB interacted with year.

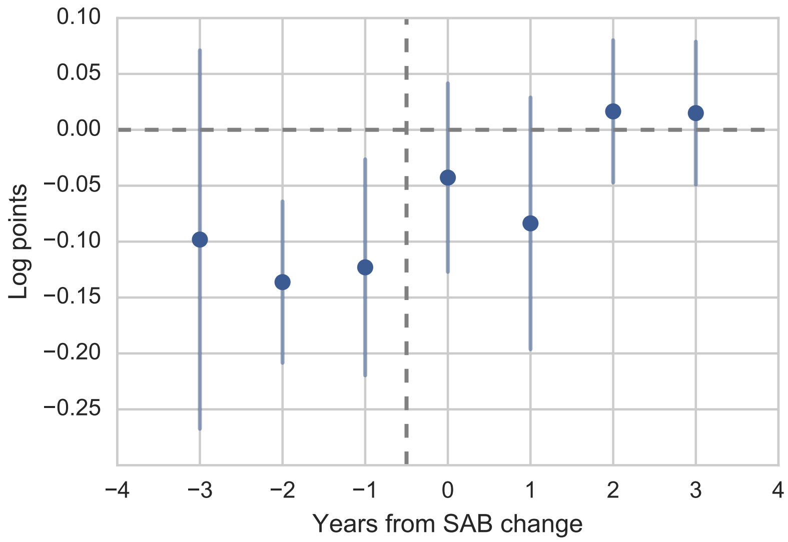


Figure 3: Price adjustment

The figure display estimates of λ^i from Equation (10). Dwelling type interacted with square meters are included as controls. Vertical bars represent 95-percent confidence interval. Standard errors are clustered by arrival SAB.

We continue to define the dummy such that it takes the value of negative one if the 'treatment" is negative, i.e. if the change in boundaries entails a negative change in school-SES. As controls, we include three-way interactions between year of sale, dwelling type and square feet. The estimation results are displayed in panel 3a. Column one displays the results without any controls. Prior to the inclusion into the SAB, the gap was approximately seven log points. Including controls, the effect jumps to ten points in column 2. This may reflect the larger share of apartments in the reassigned group compared to the control group. The estimate remains essentially unchanged when we include neighborhood controls as seen in column 3. The parameter on the treatment dummy is close to zero and insignificant once we control for observed dwelling characteristics.

These estimates are well within the range of the BDD-estimates, which are between 3 and 16 log points depending on the inclusion of controls, see Figure 1. Thus, the results provide evidence that the BDD-estimates do not suffer from significant omitted variable bias.

To convert these differences into marginal effects we replace the pretreatment interaction with the school-SES, in the school associated with

(a) Reduced form

	(1)	(2)	(3)
T × Pre	-0.0669	-0.102**	-0.0905**
	(0.0456)	(0.0392)	(0.0332)
T	-0.105*	-0.0195	0.00773
	(0.0408)	(0.0226)	(0.0198)
Cov		X	X
Nbh			X
N	395059	191297	191297

(b) Marginal effect

	(1)	(2)	(3)
School SES	0.778*	0.626*	0.473*
	(0.325)	(0.250)	(0.212)
T	-0.0932*	-0.0254	-0.000363
	(0.0368)	(0.0216)	(0.0189)
Cov		X	X
Nbh			X
N	395059	191297	191297

Table 3: Estimates of reduced form and marginal effects

The top panel display regression results from estimation of (11). The discrete transfers are translated into marginal effects by estimating equation (12). The results of this estimation is displayed in (3b). We do not impose the restriction that households should be within 2000 meters of a border. This explains the higher observation count in column 1 of the two tables. Standard errors in parentheses are clustered by arrival SAB. $\dagger p < .1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

dwelling i in period t and run regressions of the following form:

$$p_{ist} = \beta SES_{it} + \lambda r_i + \mathbf{X}_{it}\eta + \mu_{st} + u_{ist}, \tag{12}$$

where λ captures the time constant difference between eventually re-assigned dwellings and dwellings which maintain the same association throughout. Due to the fixed effect, the only variation in SES_{it} stems from SAB changes. Depending on controls, the estimates of the marginal effect of school-SES on prices fall between 0.4 and 0.8. These results are somewhat higher than the estimates of the BDD estimation, see Appendix Table B1 for reference. The coefficients on r_i are once again very small and insignificant. This indicates that the effect of school-SES is indeed causal and that the BDD estimates do not suffer from positive bias due to unobserved characteristics of the dwelling or the local neighborhood.

6 Distributional consequences of schooling

To put the estimated effects into context, we perform a simple back-of-the-envelope calculation of the consumption, which households must forego to guaranty enrollment in a high-SES school. We begin by converting house prices to annuities. In appendix B.2 we present evidence that the marginal effect of school-SES on prices (in logs) is approximately constant regardless of the magnitude of border changes. We, therefore, parametrize prices as a log-linear function of school quality, $P = exp(\beta(q - \bar{q}) + p_0)$, where $\bar{q}$ is the average school-SES which we set to 0.5 and p_0 is the reference price of the house. Assuming a T year mortgage with annual payments and a fixed interest rate of r, we can calculate the annuity equivalent of the sales price:

$$a = \frac{r}{1-(1+r)^{-T}} \times e^{\beta\left(q-\frac{1}{2}\right)+p_0}$$

We construct a simple example where a family has the choice between two identical houses (with identical amenities), but within two different attendance boundaries. In other words, p_0 is the same for the two houses. The difference in the implied annuities between an identical house associated with schools s and s' is then given by:

$$a_{s'} - a_s = \frac{r}{1-(1+r)^{-T}} e^{-\frac{\beta}{2}} \left(e^{\beta q_{s'}} - e^{\beta q_s}\right) P_0, \tag{13}$$

where $P_0 = e^{p_0}$. We make the calculation for a single-family home in 2015 prices. Given a average price of 12.000 DKK per square meter, and a 140 square meters in an average single family home, we set P_0 equal to 1,68 million DKK. We assume a loan repayment of 30 years and an interest rate of 4 percent.We calculate the difference in annuity value from moving from the 10th to the 90th percentile of the school SES distribution, which is a move from 0.36 to 0.62 in school-SES. For β we estimated effects in the range of 0.14 and 0.63.[29] We find that the mean of all estimates is

[29]In order to assume that p_0 is the same for the two houses we only use estimates, where house controls are included. The estimate of 0.14 is retrieved from column 1 in table B1 while 0.63 is retrieved from column 2 of table 3b.

	Lower bound	Prefered estimate	Upper bound
P_0	1,680,000	1,680,000	1,680,000
r	0.04	0.04	0.04
T	30	30	30
q_{10}	0.36	0.36	0.36
q_{90}	0.62	0.62	0.62
β	0.14	0.36	0.63
$a_{90} - a_{10}$,DKK	3,494.65	9,039.65	15,701.09
$a_{90} - a_{10}$,USD	513.71	1,328.83	2,308.06

Table 4: Calculation of costs

The table displays assumed values and the calculation of the lower and upper bound of the difference in annuities between two identical houses in SABs with school SES in the 20th and 80th percentile. The exchange rate from DKK to USD is 0.147.

0.36 which we regard as our preferred estimate. We calculate annuity payments from this estimate of β as well as an upper and a lower bound of the cost difference.

The calculated bounds are displayed in Table 4. We calculate a lower bound of 514 USD per year while the upper bound is 2,308 USD. Our preferred estimate is a yearly expense of 1,329 USD per year. For perspective, we calculate bounds as shares of income for each percentile in the income distribution of Danish families in 2015. We include families, where at least one adult is between 25 and 35 years old, and where there is at least one child living at home. We use disposable income after taxes and transfers and subtract the annuity value of P_0.[30] The calculated shares are therefore the share of yearly consumption that household would have to forego to buy a house associated with a school with socioeconomically strong students compared to the school with more disadvantaged students. Figure 4 displays the result, where the black lines represent the bounds and the blue solid line represents our preferred estimate.[31] For

[30]Technically we use the disposable income variable, DISPON, where we add back the rental value of property. We exclude the top 0.1 percent of the household income distribution. We subtract the annuity value of a mortgage of 1,680,000 DKK which equals 97155 DKK. We exclude households with less than 10,000 DKK disposable after subtracting the annuity, which amounts to one percent of households.

[31]The muted lines represent the shares calculated from all estimates of β. Dashed lines represent Difference-in-Difference estimates while the solid lines are from the cross-sectional BDD estimations.

the households with the highest income the costs are negligible, but for the lower income households, the cost may represent a substantial decrease in disposable income. The household at the 10th percentile will have to give up seven to nine percent of their disposable income. The cost of access to high-SES public schools can, therefore, make up a sizable budget share for low-income families.[32]

Though these calculations are subject to assumptions, they elucidate an important hindrance in ensuring equality of opportunity. Even in a welfare state with extensive transfers, poorer households need to give up a substantially larger share of their disposable income to gain access to the same educational services as high-income households. Whether these schools, with stronger peers, would be better for low-income children is a question we do not answer here. In Bjerre-Nielsen and Gandil (2018b) we find supportive evidence that stronger peers are important for low-SES children. This suggests that the cost of housing may be a blockade to achieving the full educational potential of low-SES children.

As long as children are allocated to schools via geographical zoning, we are likely to observe such patterns. A strategy to combat inequality in access is to apply more flexible admission criteria. An example of this is the results found in Machin and Salvanes (2016). The authors show that loosening the geographically defined admission criteria to high schools decreased the capitalization of school quality. If the government wants to increase equality of opportunity, this may be a possible policy. However, increasing the degree of choice may not be sufficient. If the admission systems are complicated, sophisticated households may be able to exploit the system, leaving the less sophisticated households behind. Bjerre-Nielsen and Gandil (2018a) show that this is one of the primary ways high-SES

[32]The calculation is of course simplified. We assume that everybody has access to loans at the same interest rate. Furthermore, the fixed costs of obtaining the loan do not enter into the calculation. We also assume that households may get the loan in the first place, which is not necessarily the case. All these assumptions will tend to underestimate the true inequality in access to high-SES primary schools. As we have only exploited variation in prices within municipalities, we have essentially identified after-tax valuations. The back-of-the-envelope calculation therefore implicitly assumes that the two schools are within the same municipality. In 2015, the municipal taxes ranged from 22.5 to 27.8 percent of income, with 50 percent of municipalities within 24.9 and 25.8 percent. The variation in municipal taxes is therefore not great.

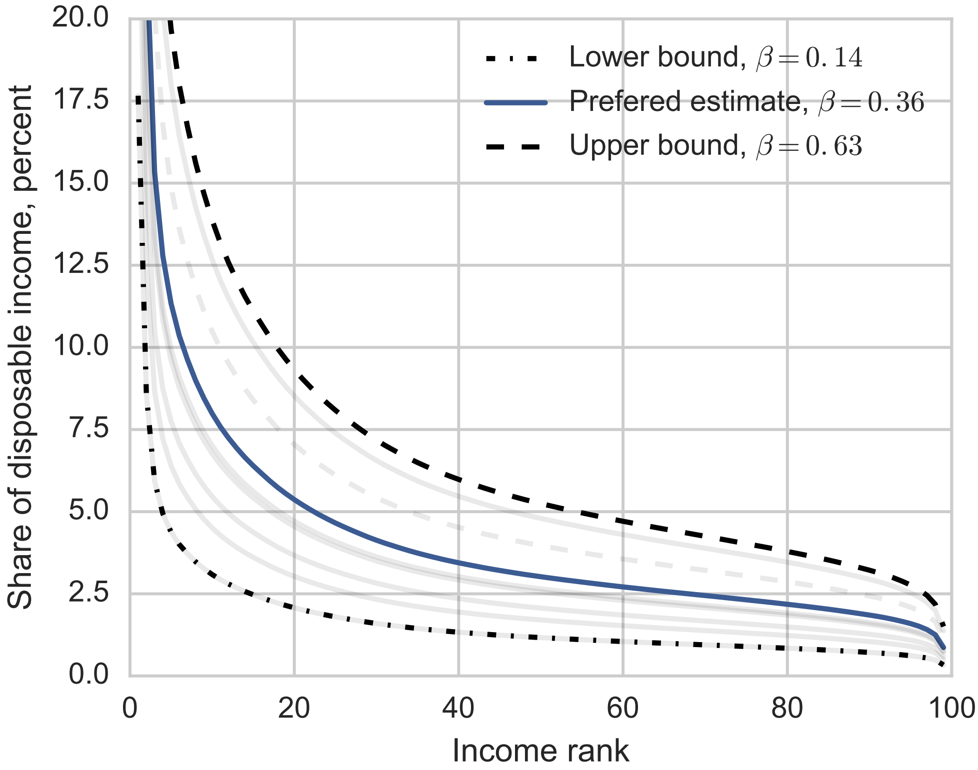


Figure 4: Annuity cost as share of income along the distribution of household income

The calculations are based on the assumptions in Table 4 and the distribution of household disposable income for Danish families in 2015. We restrict the sample to households where at least one of the adult members are between 25 and 35 years old and where there is at least one child. We use the variable DISPON as income concept and subtract the rental value of housing, calculated by Statistics Denmark. We further exclude the top 0.1 percent of the income distribution. We subtract the annuity value of P_0=1,680.000 DKK. We exclude all households with less than 10,000 DKK disposable after the annuity is subtracted, amounting to the lowest one percent. We calculate the curves for all estimates of β where house controls are included. Dashed lines represent estimates are from the Difference-i-Difference approach while solid lines represent estimates from the Boundary-Discontinuity approach.

families avoid low-SES peers in primary schools in Denmark. Loosening the admission criteria must, therefore, be done in a transparent way, such that all households may be able to navigate the process. Nevertheless, even if such a policy is implemented, the local nature of educational services will continue to create inequality to some degree if neighborhoods are unequal. Thus, without housing policy, a government is limited in its ability to ensure equal access to education and thereby possibly equality of opportunity.

7 Conclusion

In this paper, we have estimated the sensitivity of house prices to school characteristics. Using both a boundary discontinuity design and changes in school attendance boundaries. we find that prices rise with a socioeconomic index. We find little effect from test scores. We have shown that the implied price differentials between socially strong and weak schools are sizable. Low-income households may, therefore, have to give up a substantial share of consumption to buy their way into socially strong schools. Insofar as strong peers improve child outcomes, the results show that public provision of free education may be unable to ensure equality of opportunity, when children are allocated to schools according to their residential location.

References

Angrist, J. D., Pischke, J.-S. 2008. Mostly Harmless Econometrics: An empiricist's companion. Princeton university press.

Bayer, P., Ferreira, F., McMillan, R. 2007. A Unified Framework for Measuring Preferences for Schools and Neighborhoods. Journal of Political Economy, 115, 588–638.

Benabou, R. 1993. Workings of a city: location, education, and production. The Quarterly Journal of Economics, 108, 619–652.

Benabou, R. 1994. Human capital, inequality, and growth: A local perspective. European Economic Review, 38, 817–826.

Bjerre-Nielsen, A., Gandil, M. H. 2018c. Privacy in spatial data with high resolution and time invariance.

Bjerre-Nielsen, A., Gandil, M. H. 2018a. Defying attendance boundary policies and the limits to combating school segregation.

Bjerre-Nielsen, A., Gandil, M. H. 2018b. Do peers matter? only if you need them (and meet them).

Black, S. E. 1999. Do Better Schools Matter? Parental Valuation of Elementary Education. The Quarterly Journal of Economics, 114, 577–599.

Black, S. E., Machin, S. 2011. Housing Valuations of School Performance. 3, Elsevier B.V. 1st edition, 485–519.

Bogart, W. T., Cromwell, B. A. 2000. How much is a neighborhood school worth? Journal of urban Economics, 47, 280–305.

Durlauf, S. N. 1996. A theory of persistent income inequality. Journal of Economic Growth, 1, 75–93.

Epple, D., Romano, R. E. 1998. Competition between Private and Public Schools, Vouchers, and Peer-Group Effects. The American Economic Review, 88, 33–62.

Fack, G., Grenet, J. 2010. When do better schools raise housing prices? Evidence from Paris public and private schools. Journal of Public Economics, 94, 59–77.

Figlio, D. N., Lucas, M. E. 2004. What's in a grade? school report cards and the housing market. American economic review, 94, 591–604.

Fiva, J. H., Kirkebøen, L. J. 2011. Information shocks and the dynamics of the housing market. The Scandinavian Journal of Economics, 113, 525–552.

Gibbons, S., Machin, S., Silva, O. 2013. Valuing school quality using boundary discontinuities. Journal of Urban Economics, 75, 15–28.

Imbens, G. W., Lemieux, T. 2008. Regression discontinuity designs: A guide to practice. Journal of econometrics, 142, 615–635.

Imberman, S. A., Lovenheim, M. F. 2016. Does the market value value-added? evidence from housing prices after a public release of school and teacher value-added. Journal of Urban Economics, 91, 104–121.

Kain, J. F., Quigley, J. M. 1970. Measuring the value of housing quality. Journal of the American statistical association, 65, 532–548.

Kane, T. J., Riegg, S. K., Staiger, D. O. 2006. School quality, neighborhoods, and housing prices. American Law and Economics Review, 8, 183–212.

Kane, T. J., Staiger, D. O., Samms, G. 2003. School accountability ratings and housing values. Brookings-Wharton papers on urban Affairs, 83–137.

Machin, S., Salvanes, K. G. 2016. Valuing School Quality via a School Choice Reform. Scandinavian Journal of Economics, 118, 3–24.

Ries, J., Somerville, T. 2010. School quality and residential property values: evidence from vancouver rezoning. The Review of Economics and Statistics, 92, 928–944.

Rosen, S. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy, 82, 34–55.

Tiebout, C. M. 1956. A Pure Theory of Local Expenditures. Journal of Political Economy, 64, 416–424.

A Additional descriptive statistics

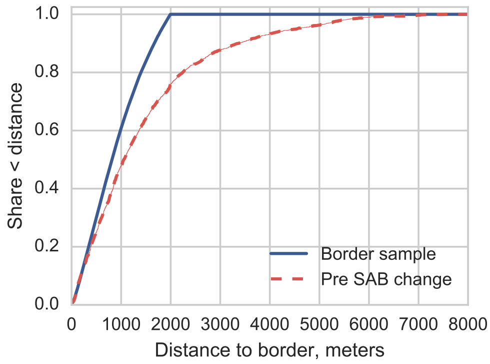


Figure A.1: CDF of distances

The figure displays the cumulative distribution functions for distances to border. For the boundary discontinuity sample distance is censored at 2 kilometres. For those addresses shifted, we only report the distance prior to the change and do not cap the distance. The border used is the border of the district to which the address is eventually transferred.

	(1) School SES	Non_Western share	GPA
School SES	1		
Non-Western share	-0.709***	1	
GPA	0.665***	-0.534***	1

Table A1: Covariance of school characteristics

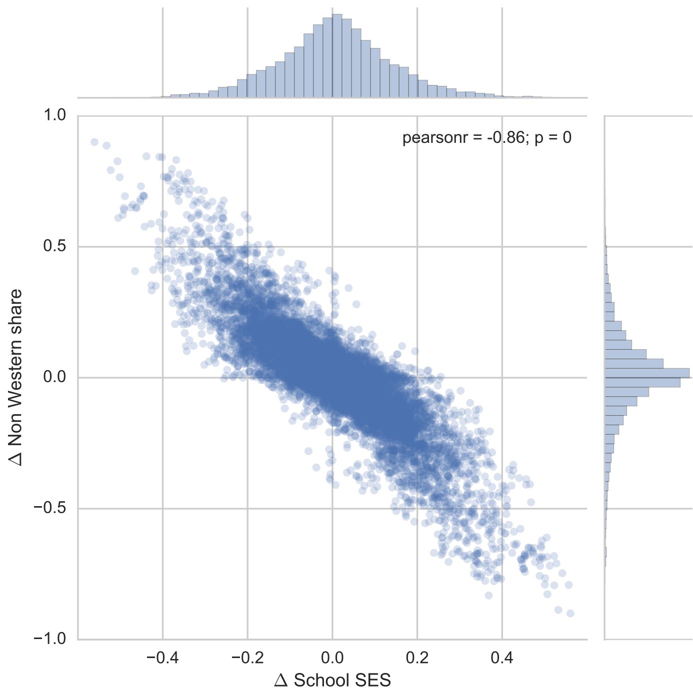
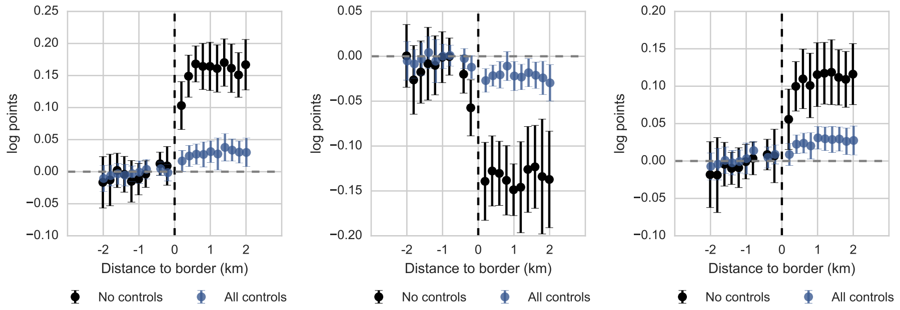


Figure A.2: Joint distribution of border differences in school SES and non-Western share

The figure displays the joint border difference distribution in school SES and non-Western share. Each dot represent a side of the border in a given year. Only border sides with more than ten observations are shown. If all borders were shown, the distribution would be symmetric around zero.

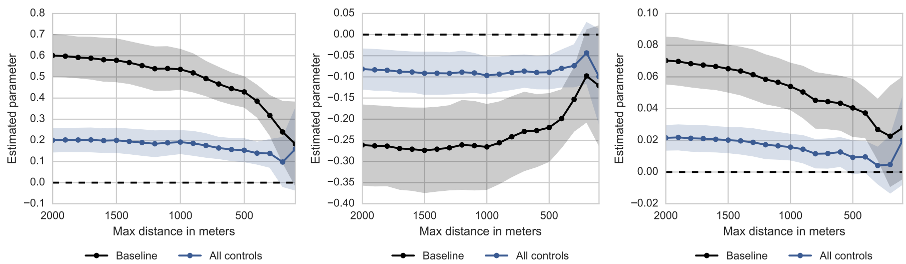
B Static analysis



(a) School SES (b) Non-Western share (c) Average GPA

Figure B.1: Simple Boundary Discontinuity Design with confidence intervals

The graphs show the results of a Boundary Discontinuity Design with one school characteristic at a time. We run regressions of discretized distances to the border, where negative distances signify that the address belongs to the side of the border with the lowest value of the measure in question. We include a border-year fixed effect to control for level differences shared by both sides of the border. In black we present the parameters on the binned distance dummies with no additional controls beside the fixed effect. The parameters in blue are from an estimation where we include hyper-local neighborhoods and square meters (including all controls squared). The results are normalized at the 400-meter distance a the left side of the border. Standard error are clustered at municipal level.



(a) School SES (b) Non-Western share (c) GPA

Figure B.2: Importance of boundaries

The figures present estimations of (9) with different restrictions on distance to the border. Going left to right the restriction becomes more restrictive.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
School SES	0.316***	0.203***	0.139***							0.572***	0.305**	0.246**
	(0.0498)	(0.0480)	(0.0345)							(0.120)	(0.105)	(0.0740)
School NW-share				-0.151***	-0.112**	-0.0733**				0.222*	0.0749	0.0561
				(0.0386)	(0.0342)	(0.0253)				(0.0895)	(0.0670)	(0.0474)
School GPA							0.0265**	0.0129	0.00432	-0.00268	-0.00661	-0.0119
							(0.00983)	(0.00879)	(0.00592)	(0.0105)	(0.0103)	(0.00651)
N: SES		1.560*	0.395		1.583*	0.412		1.535*	0.430		1.480*	0.384
		(0.649)	(0.556)		(0.659)	(0.559)		(0.681)	(0.572)		(0.665)	(0.573)
N: SES squared		-0.273	0.341		-0.281	0.334		-0.271	0.301		-0.231	0.335
		(0.543)	(0.467)		(0.554)	(0.470)		(0.565)	(0.480)		(0.550)	(0.481)
N: NW-share		0.0558	-0.120		0.0563	-0.121		0.0172	-0.157		0.0238	-0.150
		(0.123)	(0.109)		(0.124)	(0.109)		(0.123)	(0.112)		(0.123)	(0.112)
N: NW-share squared		0.149	0.281		0.161	0.290		0.251	0.375		0.232	0.359
		(0.387)	(0.287)		(0.386)	(0.286)		(0.384)	(0.284)		(0.390)	(0.289)
N: Long-cycle Educ (share)		0.516**	0.462**		0.534**	0.475**		0.536*	0.475**		0.522*	0.464**
		(0.184)	(0.142)		(0.185)	(0.144)		(0.207)	(0.168)		(0.204)	(0.164)
N: Long-cycle Educ (share) squared		0.173	-0.475*		0.143	-0.495*		0.119	-0.482		0.131	-0.473
		(0.299)	(0.226)		(0.300)	(0.229)		(0.344)	(0.253)		(0.337)	(0.247)
N: Employment		-1.349*	-0.585		-1.370*	-0.600		-1.176*	-0.477		-1.135	-0.443
		(0.552)	(0.429)		(0.553)	(0.425)		(0.582)	(0.447)		(0.575)	(0.449)
N: Employment squared		0.487	0.183		0.497	0.191		0.381	0.115		0.356	0.0951
		(0.337)	(0.285)		(0.338)	(0.282)		(0.357)	(0.298)		(0.354)	(0.301)
Single Family Home $\times M^2$			0.00302***			0.00302***			0.00299***			0.00299***
			(0.000272)			(0.000271)			(0.000277)			(0.000278)
Terraced house $\times M^2$			0.00509***			0.00509***			0.00508***			0.00508***
			(0.000258)			(0.000259)			(0.000277)			(0.000275)
Apartment $\times M^2$			0.00886***			0.00886***			0.00887***			0.00887***
			(0.000298)			(0.000298)			(0.000309)			(0.000310)
N	56374	56374	56374	56374	56374	56374	52872	52872	52872	52872	52872	52872

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B1: Full regression results

This table present full regression results for estimations of log sales prices on school, house and neighborhood characteristics. The observations are limited to observations within 300m of the boundary. All models are estimated with housetype-border-year fixed effects. Standard errors are clustered at municipal level regardless of year. The prefix "N:" signifies identifies variables calculated as hyperlocal averages.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
GPA	0.0349^{***}	0.0234^{***}	0.0135^{*}	0.0158^{***}	$0.00932^{\dagger}$	0.00432	0.00179	-0.00148	-0.00898	-0.00177	-0.00482	$-0.0119^{\dagger}$
	(0.00555)	(0.00661)	(0.00606)	(0.00418)	(0.00542)	(0.00592)	(0.00603)	(0.00718)	(0.00646)	(0.00509)	(0.00649)	(0.00651)
School SES							0.614^{***}	0.429^{***}	0.405^{***}	0.286^{**}	0.180^{*}	0.246^{**}
							(0.115)	(0.114)	(0.0839)	(0.0928)	(0.0858)	(0.0740)
Non-Western share							0.225^{**}	$0.132^{\dagger}$	0.138^{*}	0.0780	0.0151	0.0561
							(0.0800)	(0.0737)	(0.0549)	(0.0653)	(0.0563)	(0.0474)
Boundary distance	<1000m	<500m	<300m	<1000m	<500m	<300m	<1000m	<500m	<300m	<1000m	<500m	<300m
House controls	X	X	X	X	X	X	X	X	X	X	X	X
Neighborhood controls				X	X	X				X	X	X
N	193537	94495	52872	193537	94495	52872	193537	94495	52872	193537	94495	52872

Standard errors in parentheses
† $p < .1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table B2: Marginal GPA effect estimates

The models present estimates from an OLS regression of house prices on school characteristics and a border-type-year fixed effects. Standard errors, clustered at municipal level, are presented below the estimates. We estimate the model under three levels of boundary distances: <1000m; <500m, and; <300m. In some models neighborhood and other school characteristics are included.

C Hedonic regressions and the importance of schools

The formulation OF local neighborhood characteristics as bad controls is the mirror image of the argument presented by Bayer et al. (2007) to estimate preferences for neighbors. In this paper the authors argue that if sorting occurs on aN observable variable then by controlling for this variable one can causally estimate the weight put on other characteristics *over and above* the influence on school characteristics. By conditioning on the boundary fixed effect *and* school quality, we would in line with this argument be able to estimate preferences for local neighborhood characteristics. In this case sorting is seen as a function of school characteristics and not the other way around.

As we initially introduced the hyperlocal neighborhoods merely as controls to ensure exogenous variation in school characteristics we regard the interpretation as susceptible to omitted variable bias. We do not know if the exact characteristics are the characteristics that house buyers care about or merely correlate. We however still find it instructive to inspect how the inclusion of the boundary fixed-effect affects estimates of the neighborhood characteristics. To do this, we estimate hedonic regressions with neighborhood variables and house-level variables, with and without a municipality fixed effect and a boundary-year fixed effect. For the regressions without fixed effects we include dummies for years to remove shared time trends.[33] When we interpret the coefficients on neighborhood characteristics we essentially "flip" our approach compared to the main text; the neighborhood variables are now of interest and the school characteristics serve as controls. We therefore include school characteristcs jointly and abstain from interpreting them.

Column 1 of Table C1 shows the result of a simple hedonic regression. As expected local SES is positively associated with higher prices. Somewhat surprisingly, the non-Western share is significantly positive. However, once we introduce the municipality fixed effects the parameter on the local non-Western share switches signs. The strong reversal

[33]These trends are entirely subsumed into the border-year fixed effects.

most likely reflect urbanization. Non-Western immigrants and descendants tend to cluster in bigger cities where the price level is in general higher. Once we introduce the fixed effects, we are only using within-municipality variation, and the effect from the degree of urbanization therefore disappears. The valuation of neighborhood SES drops by two thirds when we include the municipality fixed effect, implying that different social classes tend to cluster in different municipalities. When we proceed to include border-year fixed effects, the valuations fall further. The parameter on neighborhood-SES is now 0.7, which is less than a fourth of the naive estimate of three in column 1. The parameter on the non-Western share in the neighborhood falls relative to column 2 but maintains the sign. One can interpret this as evidence that households sort on unobservables, even within municipalities. By introducing the fixed effect we control for unobserved amenities and the result is lower values on neighbors than a naive regression would imply, also when municipalities are controlled for.

In column 4 in Table C1 we introduce school level variables and once again omit the fixed effects. Comparing the estimates to column one we see that the neighborhood variables once again move towards zero. The non-Western share is essentially zero. Reintroducing the municipality fixed effect however, the non-Western share rises in magnitude to -0.2 but drops to -0.07 again once we include the border-year fixed effect. The latter estimate is very close to the parameter of column 3 where we included border-year fixed effect but left out school-level characteristics. This conclusion holds for neighborhood-SES as well: Once we control for boundary fixed effects, the parameters on neighborhood characteristics are not sensitive to the inclusion of school characteristics.

The importance of the boundary fixed effects found in Table C1 mirror the findings of Bayer et al. (2007) closely. Taking these results at face value, they imply that traditional estimates of the importance of neighbors are biased due to unobserved amenities. We however once again stress that we regard these results as under-identified and thus are not too comfortable believing in the exact estimates. Nonetheless, these results can serve as validation in the hedonic pricing literature.

	(1)	(2)	(3)	(4)	(5)	(6)
N: SES	3.015***	1.128***	0.738***	1.600***	0.993***	0.722***
	(0.194)	(0.0908)	(0.0487)	(0.103)	(0.0719)	(0.0504)
N: NW-share	0.843***	-0.223*	-0.0634	-0.0416	-0.286**	-0.0626
	(0.176)	(0.0980)	(0.0454)	(0.152)	(0.0860)	(0.0471)
H: Terraced house	-0.333***	-0.344***	-0.329***	-0.348***	-0.346***	-0.329***
	(0.0387)	(0.0485)	(0.0446)	(0.0353)	(0.0444)	(0.0443)
H: Apartment	-0.712***	-0.870***	-0.951***	-0.788***	-0.870***	-0.950***
	(0.0646)	(0.0364)	(0.0398)	(0.0589)	(0.0384)	(0.0398)
H: Single Family Home $\times M^2$	0.00273***	0.00326***	0.00309***	0.00279***	0.00321***	0.00309***
	(0.000230)	(0.000217)	(0.000267)	(0.000221)	(0.000225)	(0.000269)
H: Terraced house $\times M^2$	0.00543***	0.00552***	0.00513***	0.00534***	0.00550***	0.00513***
	(0.000347)	(0.000270)	(0.000303)	(0.000308)	(0.000240)	(0.000299)
H: Apartment $\times M^2$	0.00937***	0.00965***	0.00912***	0.00963***	0.00962***	0.00912***
	(0.000470)	(0.000353)	(0.000360)	(0.000401)	(0.000350)	(0.000360)
S: SES				2.065***	0.438***	0.263**
				(0.116)	(0.121)	(0.0804)
S: NW-share				1.322***	0.293**	0.0709
				(0.0904)	(0.105)	(0.0478)
S: GPA				-0.0448***	0.0100	-0.00963
				(0.0111)	(0.00597)	(0.00646)
Year dummies	X	X		X	X	
Municipality FE		X			X	
Border-year FE			X			X
N	52872	52872	52872	52872	52872	52872

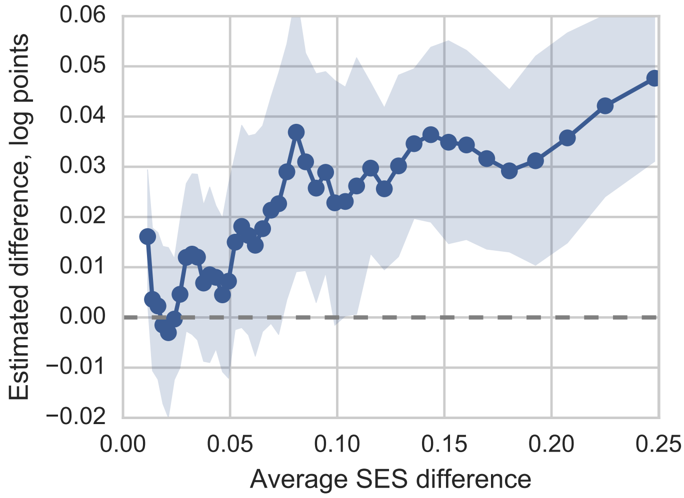
Table C1: Hedonic regressions with and without border-year fixed effects

The models present estimates from an OLS regression of log house prices on neighborhood characteristics ("N"), house characteristics ("H") and school characteristics ("S"). Model 2 and 4 include border-year fixed effects. Standard errors, clustered at municipal level, are presented below the estimates. Only houses within 300 meters of the border enter the regression. Farmhouse is the reference category.

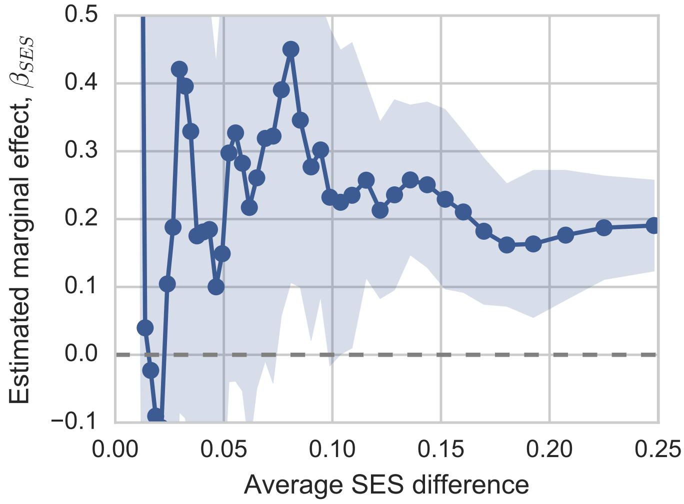
D BDD: effect heterogeneity

One may worry that the effects estimated so far reflect effect heterogeneity. To investigate this we rank the border-year combinations according to the absolute difference between the two neighboring schools. We then run our model with full controls for neighborhood and house characteristics in a running 10 percent sample window. The average difference in log prices as a function of the difference in school-SES is seen in Figure D.1. Figure D.1b we present the estimated differences as a function of the mean difference in school SES on the x-axis. We see that for small differences in school-SES there is no difference in the prices. However, once differences in school-SES exceed 5 points the relationship becomes approximately linear when. This implies that the marginal effect is approximately constant. This can be seen from Figure D.1b where we have calculated the marginal effect. The marginal effect of 0.14)found in column 3 of Table 2) is the average of an approximate (unstable) zero effect for low differences and a stable effect of around 0.2 for SES differences higher than 5 points.[34] Due to the very stable relation for SES-differences for this large region, we conclude that we do not observe a heterogeneity in the price responses according to the level of "treatment", i.e. the difference in school-SES at the border. With constant marginal effects, we can regard house prices as a log linear function of school-SES.

[34]The very unstable relationship in the bottom of the "difference distribution" can be rationalized by once again viewing the model as a Wald-estimator., where a reduced form estimate is rescaled by the first stage to obtain a marginal effect. By construction, the samples in the left most part of the figures have small differences between school SES across the border. This difference constitutes the first stage of an IV. This implies that tiny deviations in the reduced form regression, the nominator of the Wald estimator, will cause the the ratio between the reduced form and the first stage to explode in magnitude.



(a) Difference at border



(b) Marginal effect

Figure D.1: Simple Boundary Discontinuity Design

The graphs show the results of a Boundary Discontinuity Design with SES as a measure of schools. In Figure D.1a we regress log prices on a dummy for being on the right side of a district border as well as controls for house type interacted with square meters as well as hyperlocal neighborhood measures (and squares). We include a border-type-year fixed effect to control for level differences shared by the same types of houses on both sides of the border in a given year. We compute a parameter for a sliding 10 percent window according to the ranked border regions. In Figure D.1b we compute the marginal effect by exchanging the right side dummy for the level of school SES. 95-percent confidence bands are displayed as the shaded area. Standard errors are clustered at municipal level.

Chapter 4

Do peers matter? Only if you need them (and meet them)

Do peers matter? Only if you need them (and meet them)

Andreas Bjerre-Nielsen* & Mikkel Høst Gandil†

Abstract

Educational policies are often motivated by a desire to increase equality of opportunity and social mobility. However, the empirical findings on what actually works are often highly contested. This paper investigates one tool available to policy makers: the mechanism whereby students are allocated to schools. Using administrative changes in school attendance boundaries, we investigate the causal effect of changes in the expected socioeconomic characteristics of cohorts on student outcomes. We find weak overall effects. However, we show that this null result is driven by the combination of low compliance and effect heterogeneity. We find a strong, positive effect of stronger intended peers for disadvantaged children who are likely to enroll in the intended school. We document a tightly estimated zero effect for children with strong socioeconomic backgrounds. This non-linearity in effects suggests that associational redistribution may increase overall efficiency by improving outcomes for disadvantaged children at little cost to other children.

*University of Copenhagen, andreas.bjerre-nielsen@econ.ku.dk

†University of Copenhagen and the Economic Council of the Labour Movement, mga@econ.ku.dk

1 Introduction

Much economic research has documented large discrepancies in outcomes of individuals dependent on place of residence in childhood or school attendance. Chetty et al. (2014) document large discrepancies both in inequality and mobility across the US as well as within much more narrowly defined regions. In addition, Graham (2018) (among many others) documents strong correlations between outcomes and socioeconomic composition of neighborhoods. Thus, to borrow a phrase from Graham (2018), "place matters". In this paper, we look at one possible explanation for such correlations, the importance of schools.

Schools play two fundamental roles in the childhood human capital formation: firstly, they provide educational inputs such as teachers, textbooks and other educational resources. Secondly, they provide a context wherein children interact. If child outcomes are affected by other children, often referred to as peer effects, then the student composition in schools may be an important part of the explanation for the variation in performance between schools. However, educational inputs and peer composition may interact in highly non-linear ways, making the educational production function notoriously difficult to identify. It is, in other words, difficult to identify whether exposure to better peers generate better outcomes or merely correlate with unobserved school inputs. Even so, if these features continue to correlate, it may matter little for policy purposes. By manipulating one characteristic, policymakers effectively manipulate other characteristics as well.

A more pressing concern is unobserved household characteristics. If unobserved parental investments correlate with peer composition we may attribute variance to schools while neglecting the role of parents. Thus, any assessment of the effectiveness of schools must handle household sorting across neighborhoods and schools.

To identify the importance of schools, one should ideally randomly allocate children to schools. This paper investigates a quasi-experiment approximating such an experiment in Danish primary schools. The Danish system allocates children to primary schools through geographically defined attendance boundaries. Parents can therefore implicitly choose the school for their child by moving within the attendance boundary. However, these boundaries change over time, thereby allocating children to different public schools. If these changes are unexpected, then households are unable to sort in the short term. Families who thought their children would attend the same school can therefore now act as control groups for each other.

We measure the change in school as a change in expected socioeconomic composition. The effect we seek is, therefore, best thought of as an intention to treat effect (ITT). We find little overall effect of a change in intended student composition on test scores. However, we document a social gradient. We find precisely estimated zero effects of exposure to stronger peers for children of high socioeconomic status. The effect is positive for children in families with low socioeconomic status.

In Denmark, households have a degree of choice and can choose to opt out of their assigned school, a feature of the Danish system documented in Bjerre-Nielsen and Gandil (2018). This is important, as a change in the mechanical student composition may have little effect if the *a priori* likelihood of attending the designated public school is low. We, therefore, expect the magnitude of ITT-effects to be a function of the likelihood to be exposed. In the spirit of Gruber and Mullainathan (2006) we construct a prediction model for the *baseline* likelihood to comply with the school assignment. To fit this model, we use data on children living within unchanged attendance boundaries. We show that the prediction model performs extremely well out-of-sample. With this model, we predict baseline compliance for the children exposed to the exogenous changes in attendance boundaries. We then interact this prediction with the mechanical change in characteristics and document overall small and insignificant effect. However, when we interact socioeconomic status with the baseline compliance, we find that the effect for disadvantaged children is almost solely driven by children, who are likely to comply in the first place. For these children, we find large and positive effects.

Our method is not informative about the exact mechanism, but the results indicate that resourceful households are able to compensate for any adverse changes in access to specific schools. These parents could choose to compensate by moving their children out of the public sector. Though we do find evidence of this, it cannot explain the precisely estimated zero result, as the *intended* peer composition and the *actual* peer composition of these children correlate.

As we estimate positive effects for disadvantaged children and no effects for children from strong socioeconomic background, there is scope for associational redistribution. A policy whereby one mixes children from different backgrounds is likely to increase efficiency in primary education. In other words, outcomes for vulnerable children can be improved at little cost to other children.

We would like to get closer to the mechanism of these effects but are limited by our research design. Ideally, one would use instrumental variable techniques (IV) to rescale the intention-to-treat effect and thereby obtain estimates of the parameters of the production function. We argue that this is not feasible for two reasons. Firstly, the educational production function is complex and likely have multiple parameters of interest. We however only have one valid "instrument", the attendance boundary change. Thus, we find the required exclusion restrictions implausible. Secondly, we argue that the assumption of monotonicity of treatment with respect to the instrument is not satisfied. An important way that socioeconomic strong households may compensate for an adverse shock is through opting out of the assigned schools. If households exposed to a *negative* composition change end up opting for private school, the child may end up with a *positive* change. We observe strong indications of this effect in our data and therefore limit ourselves to the ITT. Importantly, given that the government uses attendance boundaries to allocate children to schools, the ITT is the effect of interest for policymakers.

The issue with compliance is twofold. Firstly, it reduces the possibilities to obtain knowledge about the educational production function. This paper documents that taking compliance seriously limits the knowledge of peer effects we can obtain from such natural experiments. Secondly, it is a constraint for the opportunities to effectuate policies to improve outcomes in primary school by manipulating student compositions.

The paper proceeds as follows: In the next section, we briefly introduce the peer effects literature along with the methodological challenges. We then proceed to present the Danish primary school allocation system in section 3 and our econometric model in section 4. In section 5 we present the data and section 6 describes our enrollment prediction model. We present the results in section 7 and section 8 concludes.

2 Peer effects and relevant literature

How education works is one of the most studied fields in economics. The literature on estimating the educational production function is enormous and impossible to cover in this short paper. We refer to Hanushek et al. (2016) for a general introduction. We will in this section focus on the research on peer effects, in other words, how students affect each other.

Theoretically, Epple and Romano (1998) and Durlauf (1996b) have shown the importance of peer effect in explaining linkages between inequality and intergenerational mobility: If socioeconomic strong children have a positive effect on other children, well-off families have an incentive to "hoard" strong peers to the detriment of less well-off families. The authors thus underscore the importance of the nature of peer effects for the desirability of a policy seeking to affect socioeconomic compositions of schools, a policy which Durlauf (1996a) refers to as *associational redistribution*. For associational redistribution to increase overall performance and thereby the efficiency of educational production, Hoxby and Weingarth (2005) argue that one needs complementarities between own and peer ability. In the absence of these complementarities, one merely reallocates outcomes and the policy therefore solely serves a goal of redistribution. For the policy to be desirable, the policymaker must in this case be inequality-averse. These studies all motivate a careful analysis of heterogeneity in peer effects for determining the scope of associational redistribution to increase aggregate welfare.

The empirical research on peer effects has a long history dating back to at least Coleman et al. (1966). The study of peer effects is made difficult by the amount of knowledge required about assignments of peers, shared inputs and group heterogeneity, brought into focus by Manski (1993) and Angrist (2014) among others. Using network analysis concepts, these studies show that partitions of agents which can be described by a block adjacency matrix can deliver little in terms of causal identification of peer effects, often referred to as the reflection problem. A block matrix is the exact way to describe classroom interactions, where children only interact with other children in the same classroom. However, Blume et al. (2015) and Bramoullé et al. (2009) have recently shown that these worries in many cases are overstated.[1] However, even in the absence of the reflection problem, the identification of peer effects entail other challenges such as endogenous network formation and unobserved covariates.

Peer effects are often thought of as an externality and are therefore not tradeable in traditional markets, see Sacerdote (2011). The channels through which children may affect each other can be numerous. Children with high socioeconomic index (SES) may affect the learning of low-SES children by directly interacting in the classroom. However, it may also be that high SES parents are able to demand better teachers for their own child, thereby incurring a positive externality on the low SES-children. In some cases, this will not be thought of as a peer effect. In the present context, however, this is a mechanism whereby attendance boundary changes may affect the outcomes of children. In this paper, we are therefore working with an expansive definition of peer effects.

At least since Manski (1993), much of the literature has been focused on

[1] Blume et al. (2015) show that some of these worries may be overcome if the network has non-transitive triads. In other words, if agent A and B are friends, then if there exists another agent C who is a friend of A but not of B then identification of a (structural) model is possible.

linear-in-means models, where the influence of peers can be described as a linear function of average peer characteristics. These models are often given a structural interpretation, for example as a function of a game, see Blume et al. (2015) for an example. Carrell et al. (2013) and Sacerdote (2011), however, show that these types of models may not be a good representation of the true social interactions as they do not account for homophily within groups and selective interactions. Using a randomized experiment, they show that when students are in very stratified groups, they interact less with students different from their own type. Similar findings have been documented by Hoxby and Weingarth (2005) and Imberman et al. (2012). An implication of these findings is that policy aimed at increasing efficiency by allocating students to affect mean characteristics may have negligible or even negative effects in practice. Accordingly, it warrants that empirical work focus on effect heterogeneity. Hoxby and Weingarth (2005) use instrumental variable estimation (IV) to get at effect heterogeneity. They use a mechanical peer composition as an instrument for true peer composition and exploit other moments than the mean to investigate the heterogeneity.

From the short outline of the literature above it is evident that any empirical approach to estimating peer effects must take endogeneity, group assignment and heterogeneity very seriously. Before presenting our econometric approach we briefly present the Danish primary school system in the next section.

3 Primary school allocation in Denmark

In what follows, we briefly describe the institutional context and our source of variation.[2] Danish public primary schools are run by the municipalities. The usual assignment mechanism is school attendance boundaries (SAB), which associate children with schools based on their residential location. Every child has a right to be admitted to the public school to which they are

[2]This section mirrors the corresponding section in Bjerre-Nielsen and Gandil (2018).

associated. These boundaries change over time due to administrative decisions, which are usually taken in the spring before schools start in August.

Once enrolled, the child is not directly affected by changes to attendance boundaries. Hence, the boundaries are therefore only important for households at the time of enrollment. The parents have the option of enrolling the child in another public school if there is sufficient capacity. Defining sufficient capacity is a decentralized and somewhat opaque process. Furthermore, parents can choose private schools, which publicly subsidized and thus fairly cheap. The private schools are, however, often heavily oversubscribed. In Bjerre-Nielsen and Gandil (2018) we investigate the choice of these two options in detail and find that especially high-SES households avoid allocations to schools deemed undesirable through both outside options. In the short term, however, the public option completely dominates the private option as a means of avoidance.[3]

We now move on to present our econometric approach and the limits to what we can infer in the present setting.

4 Econometric model

A large literature shows that choice of residence is affected by the local socioeconomic makeup, see Baum-Snow and Lutz (2011) for an example. Because schools are local goods, the socioeconomic makeup of the school will often resemble the area in which the school is located. This is especially the case when school attendance boundaries (SAB) delineate who has a right to attend which school. The implication is that similar households may locate in the same area. This poses obstacles for identifying a causal relationship between peer composition and performance. If a low-SES child in a rich area performs better than a child with similar SES in a poor area, we cannot readily conclude that the peer group explains this variation. We simply do not

[3]This is important, as it diminishes issues of non-random attrition due to the lack of test scores for students in private school. See the last part of section 7 for sensitivity checks.

know why the two children live in different places in the first place. In other words, the non-random residential location decision makes any estimated effects of school peers susceptible to omitted variable bias.

In this paper, we will rely on natural experiments induced by administrative changes in attendance boundaries. In order for such an approach to yield unbiased results, we assume that household sorting is fully controlled for if we know the school to which the households *sought* to associate themselves before the administrative change. Assuming that households do not anticipate the boundary changes, we can control for residential sorting by including "original SAB"-year-fixed effects in our regressions.

To characterize the relocation, we need a way to summarize the variation in the peer characteristics induced by the boundary changes. We, therefore, construct a "mechanical" school characteristic in the vein of Hoxby and Weingarth (2005). We observe all children, their characteristics and addresses at age 5. Using this data, we construct our measure of mechanical school characteristics by averaging over the set of children living within the SAB to which their addresses will be associated *two years later* when the children are supposed to be enrolled in primary school. We focus on the average socioeconomic index and call this measure the *mechanical school-SES* (MSS henceforth). We describe the construction of the index in section 3. For the MSS to be perfectly correlated with the actual characteristics, we would need stable SABs and restrict households from moving, delaying school start or choosing other options than the local school. Thus, the MSS will naturally differ from the actual measure to which the child is exposed. It is, therefore, best thought of as an instrument for actual peer composition.[4]

In the language of the LATE framework, we estimate a reduced form regression of this instrument, the mechanical schools SES, directly on the outcome of interest. We perform regressions of the following kind:

[4]However, we later argue that is is not entirely clear for which characteristic the MSS is an instrument. While we use IV parlance, we do not actually use IV.

$$Y_{iss't} = \alpha X_i + \beta MSS_{s'(t-k)} + \mu_{s(t-k)} + \varepsilon_{iass't} \tag{1}$$

where $Y_{iss't}$ is the outcome of child i living at an address which belongs to district s at time $t-k$ but to district s' at time t. This outcome is assumed to be a function of child's own characteristics, X_i, which does not vary over time. We are interested in β, which is the parameter on MSS. As mentioned, we control for residential sorting by including a SAB-year fixed effect. The only variation in MSS therefore comes from administrative boundary changes, where children within the original attendance boundary are allocated to different schools and therefore have different realizations of MSS.

If we were to interpret equation (1) as a structural model it would leave little room for associational redistribution. This is because the model does not allow for complementarity between own and peer characteristics. We, therefore, estimate variations of the model with heterogeneity in own socioeconomic status. It should therefore also be clear, that we do not interpret equation (1) as reflecting a structural relationship between peer characteristics and own outcomes.

From reduced form to IV? The reduced form model in (1) amounts to the aggregate effect of the change of SAB characteristics on outcomes. We do not know how much of this effect comes from actual peer characteristics and how much comes from other factors related to the intended changes in peer composition. The natural next step would therefore be to estimate the first stage and combine this with the reduced form estimate to obtain an IV-estimate, as is done by Hoxby and Weingarth (2005). However, this approach is fraught with peril.

From the literature we have good reason to expect effect heterogeneity. In this case, an IV-regression is best interpreted as a local average treatment effect (LATE). For such an estimate to be unbiased, we need three assumptions; exogeneity of the instrument, an exclusion restriction and monotonicity in the instrument. In the present context of estimating peer effect, we have faith in the exogeneity of attendance boundary changes. We, however, doubt that the other two restrictions are fulfilled.

Firstly, if we run a first stage, where MSS is used as a predictor of actual cohort composition, and scale our reduced form estimate with this first stage, we assume that it is actual cohort-SES which explains the variation in performance induced by the boundary change. However, household-SES correlates with other socioeconomic factors such as ethnicity. As we only have one instrument, we cannot separate out factors correlated with SES. We therefore only view our SES-index as a proxy for the set of factors which might affect child outcomes. Further, we do not know whether the link between SES and performance is due to actual *interaction* or whether variation in peer composition induces different teaching strategies or other educational inputs. If this is a function of actual peer composition this is not a problem, but if it is a function of *expected* composition, then the exclusion restriction is not valid.

Even if the exclusion restriction is valid, the issue of monotonicity remains. This assumption normally receives relatively little attention but is crucial in our framework. If the residential address of a household becomes associated with a school with sufficiently low expected SES, the household may opt for an outside-option such as private school or the possibility of enrolling in other public schools with sufficient capacity. If this outside option has a higher SES than the original district, monotonicity breaks down; variation in the instrument may induce treatment of the opposite sign from what the instrument would suggest. This would imply that some groups would have negative weight in the computation of the average local treatment effect, which then seizes to be meaningful. This issue is a function of the behavioral responses of households documented in Bjerre-Nielsen and Gandil (2018). In appendix A we show direct evidence that the assumption of monotonicity is indeed invalid in our setting. In other words, even if the exclusion restriction holds, the presence of heterogeneity in treatment combined with defiance prevents us from interpreting an IV estimate as a local average treatment effect.

These two issues are sufficiently severe for us to abstain from estimating IV regressions. We therefore focus exclusively on the reduced form regression and estimate Intention to Treat effects (hereafter ITT). Importantly, from the policymakers perspective, this may not be an issue as the "instrument" is policy relevant. Regardless of the household responses, we can estimate the effect of redrawn attendance boundaries. Thus, in our setup, the reduced form regression is informative about the effect of the policy tool, which municipalities have at hand.

4.1 Identifying affected children

Having shown that there are fundamental problems with an IV approach, we need alternative strategies to understand the mechanisms for a given reduced form effect. The implication of the discussion above is that not all types of households are affected the same way by the instrument and that the behavior of the households may be non-monotonic in the MSS. If an effect goes through the actual exposure to peers, we would expect larger effects for children who comply with the assignment to primary school, compared to children who do not attend the intended school. Using the actual enrollment compliance, however, will constitute a problem of being a 'bad control' as the likelihood of compliance depends on the treatment, i.e. changes in the mechanical school-SES. Because the choice to comply is not random, conditioning on enrolling would introduce selection issues, which are not present in the baseline setup. To circumvent such issues, we take an approach inspired by Gruber and Mullainathan (2006). We seek to identify the effect on those children who are *likely* to comply by the assignment mechanism in the absence of a change to the attendance boundary, rather than those who *actually* comply.

To do this, we divide our sample into two subsamples. The children in the first sample experience changes in their school association through attendance boundary changes. This sample constitutes our main dataset and we call this the *causal analysis sample*. The other sample, the auxiliary sample, consists of children who experience no changes in the boundaries. We use this dataset to construct a prediction model for enrolling in designated school given school characteristics at age five and family background. Note, that we only claim to have exogenous variation in our causal analysis sample. Hence, we do not consider our prediction model to be causal. We now move to describe in general terms how we construct the prediction model.

4.2 Predictive modelling

We construct our model based on household and neighborhood variables and expect many of these variables to interact in highly non-linear ways. We could model this with interactions of the variables in a linear regression framework. However, a central worry would be that we construct a model that predicts well on the estimated data but with poor out-of-sample performance. To avoid this overfitting problem, we apply a *Random Forest* algorithm (RF), introduced by Breiman (2001). This algorithm relies on fitting classification trees which handle non-linear data well, see Wager and Athey (forthcoming). However, instead of fitting a single tree, RF constructs multiple trees on different subsamples of the variables and observations. The method thereby reduces the risks of overfitting. Using the multiple decision trees, RF forms a prediction based on the average classification of the estimated trees. We can interpret this average as a probability.

A drawback of RF is that it is less transparent than a linear model. Evaluating the impact of a variable on the prediction may, therefore, require some form of simulation. In order to validate the model, we use out-of-sample prediction. We construct our model on eighty percent of the data in the auxiliary dataset and use the remaining twenty percent for model evaluation. Importantly, this entire process is kept separate from our actual analysis using changes in attendance boundaries.[5]

[5]We use the implementation of RF in 'scikit-learn' using the Classification and Regression Tree algorithm (Breiman, 1984; Pedregosa et al., 2011). We estimate the Random Forest with the hyperparameters recommended by Breiman (2001). Most importantly this means

It is important to compare the prediction model to the alternative strategy of simply including covariates directly in the regression of test scores on the intended peer characteristics. This would implicitly control for compliance but would not be informative about the mechanisms. We would therefore not know whether any sensitivity to controls is caused by actual effect heterogeneity or differing levels of compliance. However, our prediction is essentially a non-linear function of socioeconomic variables, as we use these variables to construct the prediction. Thus, the use of this prediction together with controls implicitly exploit the same variation twice. We return to this issue in section 6.2.

5 Data and measurement

This section describes the choices made in structuring our data and closely mirrors Bjerre-Nielsen and Gandil (2018) as the identifying variation used in these two paper is the same.

We base our analysis on Danish registry data covering the years 2008-2015. From the registries, we obtain information on households such as child gender, ethnicity, number of adults in the household, parent income and education and, importantly, geographical location. This data is of very high quality and cover the universe of Danish children and their parents. From the CPR-vej-registry we obtain information on school attendance boundaries (SABs). The municipalities report these boundaries as sections of roads. Submission is voluntary, and for the municipalities own use in their administrative IT systems. Statistics Denmark do not verify the data accuracy. We clean the SAB-data and merge it unto the register data using the variables *kom* and *opgikom*.

We sample all 5-year old children who are observed two years later, en-

using at most a number of variables for each tree equal to the square root of the total number of variables. We estimate our models using 1000 trees to ensure minimal overfitting. The mean of the votes cast by the thousand trees is interpreted as the predicted compliance rate.

rolled in primary school. Most children should be enrolled at age 7, even if parents have chosen to postpone school by a year. For outcomes, we use test scores from low-stake tests taken in *public* primary schools, which we obtain from the registries. Tests are taking almost every year, alternating between Danish and Math in the early years. We focus primarily on tests in Danish language, taken in second grade – the earliest possible test. Within a subject, students are scored along three different dimensions. We take the mean of these three dimensions and rank within cohort. Our measure of performance, therefore, follows a uniform distribution on the unit interval. As private schools do not take the tests, the sample suffers from non-random attrition, A point to which we will return in section 7.3.

To summarize across the multidimensional socioeconomic space we construct a socioeconomic index as the first component from a principal component analysis. We rank this measure across cohorts to achieve a uniform index. The socioeconomic index is increasing in income, employment and high cycle education as expected. See the appendix of Bjerre-Nielsen and Gandil (2018) for further details. We refer to this index as *household-SES*. We use an average of the SES as a proxy for the socioeconomic student composition in public schools and for calculating the mechanical school-SES - the central measure of interest in the analysis as described in section 4. After calculating school measures, we exclude non-Western children from the sample. We elaborate on the reasons for this restriction in section 6.2.

5.1 Descriptive statistics

Table 1 presents descriptive statistics for the causal sample. The second row is the mechanical school socioeconomic index (MSS), demeaned at the "attendance boundary"-year level. This is our identifying variation. Even in the causal sample, we see that the mean is zero. Appendix Figure B.1 shows that the distribution is fairly symmetric. Household-SES is slightly higher than 0.5, which means that the households living within unstable boundaries are somewhat higher on the socioeconomic spectrum than the

	Mean	Std.	Median	N
MSS	0.550	0.114	0.552	15,648
MSS, demeaned	0.000	0.035	0.000	15,648
Household SES	0.590	0.284	0.625	15,648
Female	0.488	0.500	0.000	15,648

Table 1: Descriptive statistics for causal sample

This table describes the data excluding children of non-Western descent, as they are taken out of the regression analysis. In the second row we demean the MSS with the SAB-year fixed effect.

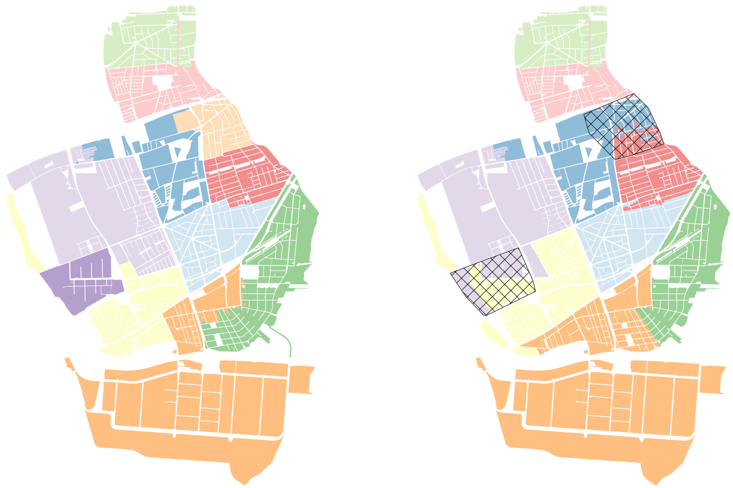
average household with children in primary school.

5.2 Empirical example

To clarify the kind of variation we exploit, we briefly illustrate a case present in our data. Figure 1 shows the school districts of the municipality of Hvidovre in the Copenhagen metropolitan region in 2011 and 2012. Between these years, two schools close down. Thus, households living within the attendance boundaries of now the closed schools unexpectedly become associated with new schools. If these households have unobservable preferences for schools, which may correlate with outcomes, we can expect these to be shared among all the households within the same original boundary. The households can therefore act as control groups for each other. By including a fixed effect for those households, which have children of the same age and thought they would send their kids to the same school, we reduce the risk of omitted variable bias caused by residential sorting.

The mechanical SES is essentially calculated as the population distribution in the map in Figure 1a with the attendance boundaries of the map in 1b superimposed.[6] Observe that besides the two school closings, other boundaries also change. With our measure of mechanical SES, we pick up this variation as well.

[6]In reality, the base map would be 2010, as we measure the MSS using the boundaries two years later, rather than one year later as the example in Figure 1.



(a) School districts, 2011 (b) School districts, 2012

Figure 1: District changes in the municipality of Hvidovre

The figures depicts the school districts in the municipality of Hvidovre in the autumn of 2011 and 2012. The hatched areas in 2012 show the convex hull of two closed schools. In order to enroll students who would live in districts of the now-closed schools, a range of other changes was made. See section **??** for a description of the district data. Some areas differ from the official documentation. These areas are mostly not populated but some measurement error occurs. The map is constructed by merging addresses on to official geo-data. In the analysis we use addresses directly to bypass mismeasurement of geographical entities.

Variations in the mechanical SES thus occur for two reasons. Firstly, some children become associated with a new school. Secondly, there are children who are continuously associated with the same schools, but where other students become associated or disassociated with that school, thus altering the expected socioeconomic composition of students. All children within changed boundaries, therefore, experience "treatment", regardless of whether the actual school assignment changes.

6 A model of enrollment

We proceed with developing a model for predicting enrollment of households into the intended school. As mentioned in section 4, we estimate the model on the auxiliary sample without exogenous variation and use the model to infer compliance in our main causal sample, where there is variation in the attendance boundaries. An illustration of the sample-splitting process is shown in figure 2.

The model forms a probability of attending the school to which the address of a child is associated based on two kinds of variables. Firstly, we record household variables such as family SES, income, type of residence, gender and ethnicity. Secondly, we use average cohort characteristics of the children within the same attendance boundary at age 5.

6.1 Evaluating the prediction model

We begin by inspecting which variables are important for the predictions in the model. Figure 3 displays the importance of the variables used. Household income rank and SES are the most important variables for determining predictions in the model. Furthermore, most of the cohort averages are more important than the remaining household level variables.

For all observations in the test sample, we predict the probability that the child enrolls in the school to which they were associated at age 5. We achieve an accuracy of 83.8 pct. in the out-of-sample prediction, which we regard as high.[7]

To visualize the fit, we bin household into SES-percentiles and take the average of the probabilities. We compare this to the mean of actual compliance for the test sample. Figure 4a shows the result of this exercise. The

[7]We decompose model accuracy by error type. Our model has a precision score of 77.8 pct. and is measured as the ratio of those who actually enroll in the intended school over those who are predicted to enroll. The rate is around 5 pct. higher than the mean rate of enrollment in the test data set of 73.5 pct. Our model has a recall accuracy of 90.7 pct. which is measured as the share predicted to enroll among those who actually enroll.

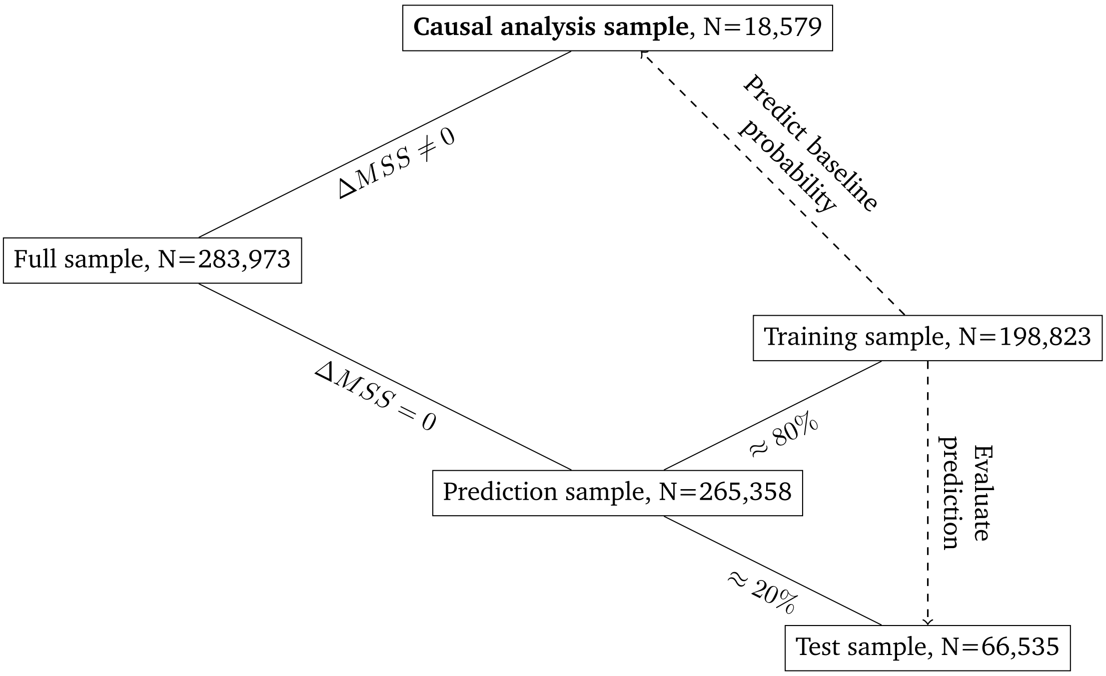


Figure 2: Definition of samples

This figure documents the sample definitions. We divide the full sample into two. The first sample, the causal analysis sample, contains those children who experience a change in the expected mechanical SES from age 5 to age 7 which imply a change in attendance boundaries. The children in the second sample, the prediction sample, experience no change in the expected mechanical SES. We further subdivide the prediction sample into two, a training sample and a test sample. We train our prediction model for assignment compliance on the training sample and evaluate the fit on the test sample. After evaluation we then use the models to construct a new variable in our causal analysis sample; a baseline probability of compliance.

black line is the actual mean of compliance in the test sample. For low SES-households compliance is in general low. We speculate that this feature is most likely due to urbanization and therefore the geographic density of schools. Approaching the middle of the SES-distribution, compliance rises and thereafter falls slightly. The functional relationship between household-SES and compliance is therefore not monotonic. Nevertheless, our prediction model is able to fully capture this, as evidenced by the blue line in Figure 4a. The apparent discontinuities are to be expected as SES is an index

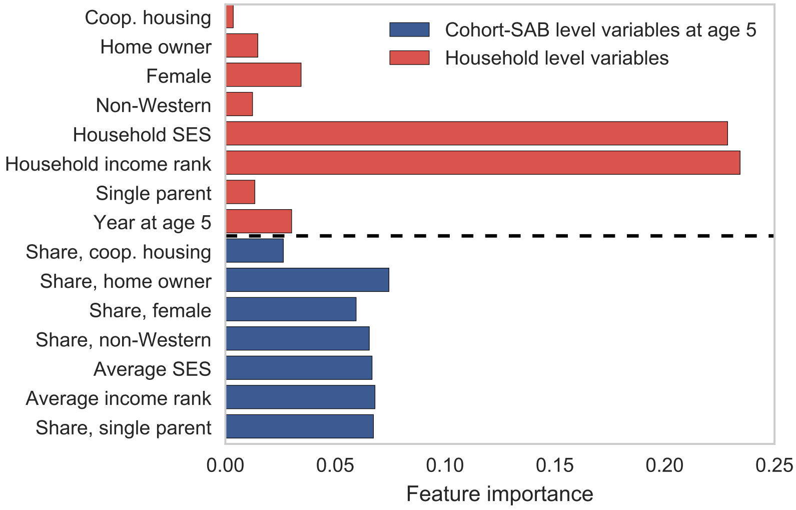
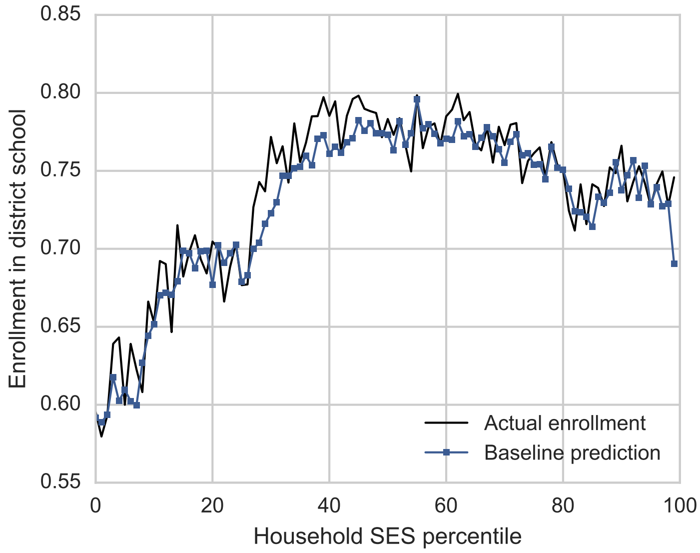


Figure 3: Feature importance in fitted prediction model

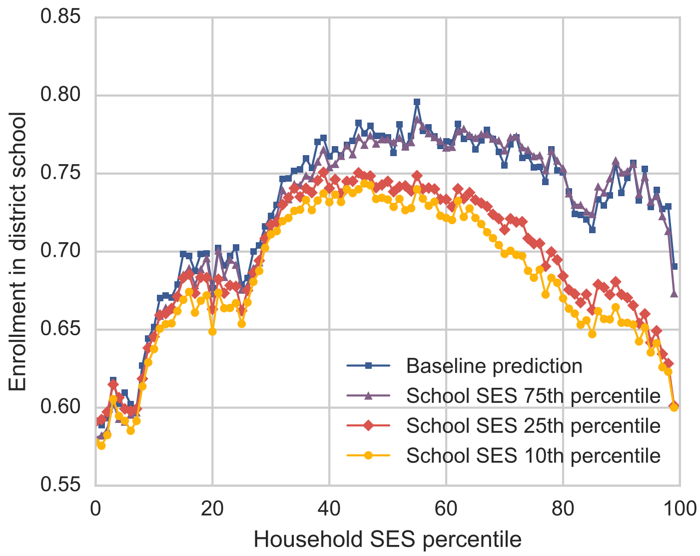
The figure presents the normalized Gini impurity measures for the features in the fitted Random Forest model. Intuitively a higher measure implies that predictions would suffer more from excluding the variable from the model. Only characteristics at age 5 are used as covariates. The measures sum to 1.

constructed from multiple variables. Most of these underlying variables are however not part of the variables used for training the model. Nevertheless, the algorithm is able to learn this process and accurately predict propensities to comply. This provides evidence that our model captures central elements of the decision process for households about whether to comply with the allocation mechanism.

To get a sense of how the model works, we can synthetically manipulate the cohort characteristics in our test sample and observe the changes in prediction. We do this in Figure 4b, where we repeat the baseline prediction from Figure 4a in blue. We first raise every household's expected school-SES to the 75th percentile of the school-SES distribution. We see almost no change in the predictions, as evidenced by the purple line. However, when we lower the school-SES to the 25th percentile we see strong heterogeneous effects (red line). While lower-SES households barely change behavior, higher SES households opt out of the assigned school. From around the



(a) Out of sample prediction



(b) Synthetic shocks to school SES

Figure 4: Performance of prediction model

The figures present the results from the fitted Random Forest model. Only characteristics at age 5 are used as covariates. Figure 4a compares out-of-sample predictions to actual behavior. For each SES-percentile in the test sample we calculate the mean of a dummy for compliance (black line) and mean predicted compliance (blue line). Figure 4b evaluates the model by changing the expected cohort SES while keeping other school and household characteristics constant.

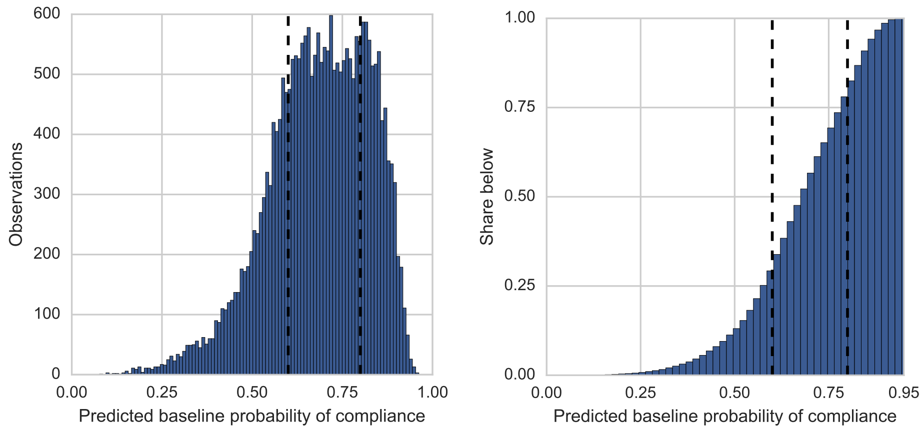
40th household percentile, the drop in compliance is increasing in household-SES. If we decrease school SES further, the magnitudes only become larger.

These results are in line with our findings documented in Bjerre-Nielsen and Gandil (2018). However, this is *solely* a predictive model for a stable environment. We do not exploit any exogenous variation and cannot claim that these effects are causal. Nevertheless, the out-of-sample predictions imbue us with confidence that the model provides a good approximation of baseline compliance in the absence of exogenous changes in attendance boundaries.

6.2 Predicting baseline propensities in causal sample

We now take our prediction model to the causal analysis sample and construct a predicted baseline probability for each observation in the data. The density of predicted baseline probability in the causal sample is displayed in Figure 5. We group the predictions into three roughly equally sized groups: below 60 percent, between 60 and 80 percent and above 80 percent. These are marked by the dashed vertical lines in Figure 5.

In our causal analysis sample, the “prediction errors” are of interest in and of themselves. They elucidate how the enrollment decision, i.e. compliance, is a fundamental hindrance for estimating peer effects. We plot the misclassification errors as a function to the variation in mechanical school-SES in Figure 6. The classification errors are divided into false positives (falsely predicted to be in the associated school) and false negatives (falsely predicted to opt out of the associated school). Reassuringly, errors are minimized when there is approximately zero change in MSS. However, as the changes in MSS become larger, so does the error. The false positive rate rises as the shock become more negative. In other words, when there is a negative shock to the average expected SES, parents tend to defy the school assignment. Conversely, the false negative rate rises as the shocks become more positive. This implies that faced with a positive shock, parents tend to opt in at a larger rate than the baseline prediction would suggest. This closely



(a) Density

(b) CDF

Figure 5: Distribution of baseline predictions in causal sample

The two figures display the density and cumulative distribution function of predicted baseline probabilities. The probabilities are out-of-sample predicitons based on the Random Forest model presented in the text. Only baseline characteristics at age five are used for the model. The vertical dashed lines represent our grouping of the predictions into three groups; below 0.6, between 0.6 and 0.8 and above 0.8. These groups correspond roughly to a third in each group as one third of the sample is below 0.63, two thirds are below 0.77.

mirrors our findings in Bjerre-Nielsen and Gandil (2018). The changes in boundaries play no role in the baseline prediction as we only use neighborhood and school variables at age 5, which is prior to variations in the boundaries. The predictions are therefore the counterfactual, the predict compliance in the case where the attendance boundaries had not changed.

The results of Figure 6 show the perils of estimating the peer effect by IV. The magnitude of the instrument is correlated with compliance rates. The non-compliance explains why we observe non-monotonicity in the instrument. In other words, when the instrument is “large”, the actual treatment may go in the opposite direction. We document this in Appendix A. Once again, and in light of these observations, only estimating the reduced-form seems like the most sensible option.

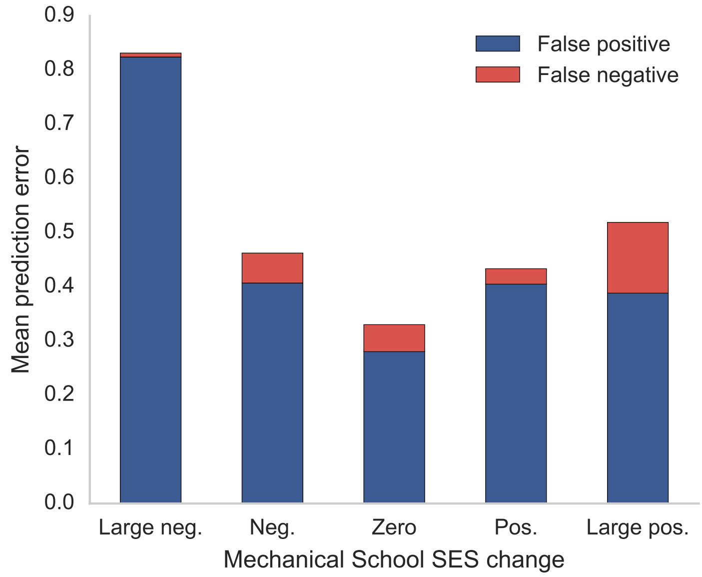


Figure 6: Prediction errors in causal sample

The role of controls and identification of likely compliance As the attendance boundaries are not stable in the causal sample, the calculated prediction is counterfactual. In other words, we are predicting the counterfactual compliance *in the absence of changes* in attendance boundaries. The nature of the prediction model implies that the probability is essentially an index of covariates. This has implications for which variables to include as controls in our regressions. In the limit, where we include covariates in a fully saturated model, the baseline prediction should have no explanatory power as it is a function of these same variables. If we excluded control variables in the main regressions and these same variables are used to form the predicted compliance, it amounts to an assumption that the variables affect test scores primarily through exposure and not through direct effects. This is a non-trivial assumption and we can not test whether it is true. I other words, we do not have sufficient variation to partial out the two mechanisms.

A poignant example of this issue is the question of including or excluding children of specific ethnicities from the sample. A child being non-western is important for predicting compliance in our model. At the same time, average

Danish test scores are significantly lower for non-Western children in the non-causal sample.[8] We found that our initial estimations were very sensitive to the inclusion of non-Western children in the sample. However, we do not claim to know whether being of non-Western descent primarily affects children through compliance (and thus exposure to peers), inherent ability or parental inputs and therefore exclude this group from the regressions. One could use this argument for other socioeconomic characteristics as well.

Due to the out-of-sample performance of our compliance predictions, we believe that the model carries weight and that the covariates may work through the propensity to comply. However, while we think we capture effects stemming from exposure, the example with non-Western children illustrates that the essential uncertainty about mechanisms remains. The results from using the predicted compliance should therefore be regarded as qualified guesses at the underlying mechanisms, rather than concrete proof that predicted exposure is a driver of effects.

7 Results

We begin by regressing test scores on SES-quartile and mechanical changes in school-SES (i.e. MSS). The estimates are found in Table 2. We see from column 1 in Table 2 that the average effect of MSS is positive but insignificant. This implies little overall effect once we control for residential sorting. This changes when we interact the mechanical school-SES with household-SES quartile. In column 2 the reference group is the lowest quartile. There is a large and significant positive effect of MSS on language test scores, as seen by the parameter on MSS. If the mechanical school-SES rises by a standard deviation (≈ 0.1), the expected test score rises by 2 percentiles. In other words, low SES children gain from potential exposure to children from more

[8]A simple linear regression of test scores on a non-Western-dummy gives a coefficient of -0.18 with a t-value of -25.56. Baring in mind that the test scores are bounded between 0 and 1, this effect is quite large.

	Danish, grade 2		*Math, grade 3*	
	(1)	(2)	(3)	(4)
MSS	0.117	0.226*	0.0534	0.127
	(0.1000)	(0.104)	(0.112)	(0.133)
MSS x SES Q2		-0.132		-0.0885
		(0.0813)		(0.0784)
MSS x SES Q3		-0.130		-0.0778
		(0.0789)		(0.0768)
MSS x SES Q4		-0.160*		-0.107
		(0.0807)		(0.0863)
N	11669	11669	10298	10298

Standard errors in parentheses
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Effect of mechanical SES changes

This table present different specifications of the model presented in equation (1). The dependent variables are test scores for Danish language in second grade and mathematics in third grade. All models include fixed effects for year-district at age five and a full set of controls including household socioeconomic ventile and gender. We cluster standard errors at the school associated with SAB at age 5. In other words, the clusters consist of all children who originally sought to sort into the same school.

advantaged backgrounds. The interaction terms are all negative and rising in magnitude. This implies that the effect becomes smaller as household-SES rises. The point estimates show that for a standard deviation in MSS, the expected test score of high-SES children rises by 0.6 percentiles. This effect represents less than a third of the effect for the low-SES children.

When we change the dependent variable to math test scores from third grade, we observe the same patterns, though the effects are in general insignificant. As the test is taken a year later, this is not surprising. We would expect the mechanical school-SES to be less important for the test variable when households have a longer time to adjust to the change. In an IV setup, this smaller reduced form effect would be counteracted by a smaller first stage, which would adjust the IV estimate upwards. As explained, we do not deem this approach suitable due to the lack of a valid exclusion restriction and non-monotonicity of the instrument.

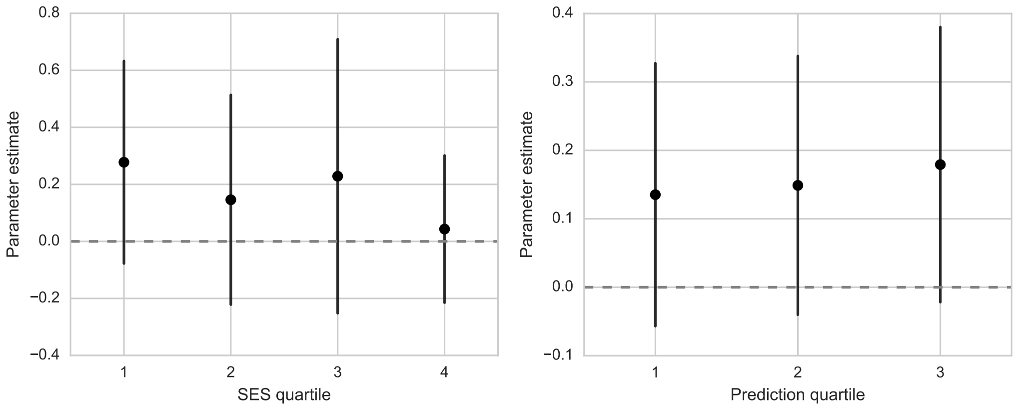
7.1 Predicted compliance

We now proceed to include our compliance measure into the regressions. We fully interact the mechanical change with household-SES quartile. We include a fully interacted set of household-SES-ventile-dummies and baseline-prediction-ventiles as well as interactions to control for nonlinearity, which might confound our findings. Thus, the controls are more extensive than in the previous section. We continue to include the SAB-year fixed effects and cluster standard errors at SAB-level. To ease interpretation, we plot the interaction terms in Figure 7a. The parameters are all insignificant, though we stress that the number of controls greatly reduces statistical power. The effect is however still decreasing in SES as before. For the highest SES quartile, the interaction term is a precisely estimated zero.

Next, we perform the same regression, but instead of interacting the mechanical school-SES with household-SES, we interact the former with dummies for the three prediction groups described in section 6. As can be seen from Figure 7b, the effect of MSS is increasing in the probability of compliance, though all estimates are insignificant.

In Figure 8 we interact MSS with our three groups of predicted compliance and household SES quartile. We maintain SES-SAB-year fixed effects and interactions between household SES and baseline predictions. For the three highest quartiles, we find small insignificant effects. For the highest quartile, we identify a precisely estimated zero regardless of prediction group.

For the lowest quartile, however, we see that the cross between likely compliance and household-SES is important. For the subset of children with a high probability of compliance and a low socioeconomic background, we see a relatively large significant effect. For the lowest prediction group in the lowest SES-quartile, the estimates are close to zero and insignificant. In other words, we find weaker effects for those households, who we a priori do not expect to enroll their child in the assigned school. The differences between predicting groups within SES-quartile are however not statistically



(a) Split by household SES (b) Split by prediction group

Figure 7: Heterogeneity in effects of MSS on test scores

The parameters are estimate in a model with three way interactions between prediction group dummies, SES quartile dummies and MSS as a continuous measure. Two-way interactions between prediction vintiles and SES vintiles and interactions between non-Western and gender dummies are included as controls. The dependent variable is test scores in Danish language taken in second grade.

significant.

We once again stress that the baseline predicted compliance is a function of observed covariates and that the non-linearities are essential for identifying the heterogeneity stemming from compliance separate from other controls. However, with this in mind, we find the results to be a strong indication that the effects of changes in attendance boundaries are higher for those likely affected by the change.

Figure 8 demonstrates why we find little effect in the simple regression of test scores on mechanical SES. The effect is heterogeneous in SES, low SES children benefit while high SES children do not. However, this may not be the whole story. While all low SES children may benefit from stronger peers, we have no way of affirming this as they are not all exposed to the "treatment". In other words, the majority of the sample shows no effect of the changes induced by the boundary changes, either because they are insensitive to peer compositions or because they never experience a change

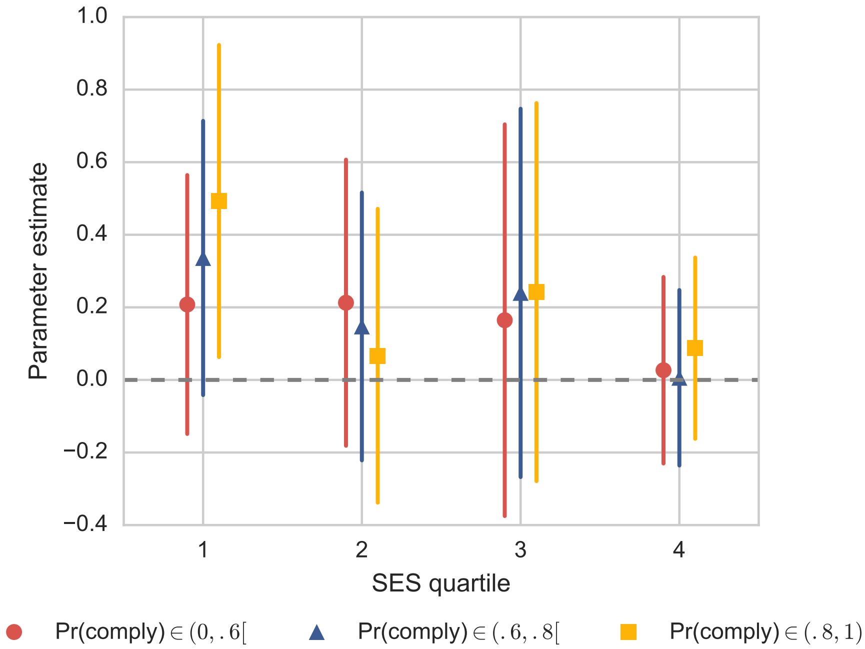


Figure 8: Heterogeneity in effects of MSS on test scores, interaction

The parameters are estimated in a model with three-way interactions between prediction group dummies, SES quartile dummies and MSS as a continuous measure. Two-way interactions between prediction ventiles and SES ventiles and interactions between and a gender dummy are included as controls. The model is estimated with a fixed effect for all SES-quartile-SAB-year combinations. The dependent variable is test scores in the Danish language taken in second grade.

likely due to non-compliance.

7.2 Exposure

If the lack of effect for high SES-groups is due to non-compliance, we should see little effects of the mechanical changes on actual cohort-SES of the children. To investigate this, we perform the same estimation as in Figure 7 but exchange the dependent variable for the average cohort-SES in the school in which the child ends up enrolling. In this regression, we have not conditioned on the availability of test scores, and we, therefore, include all chil-

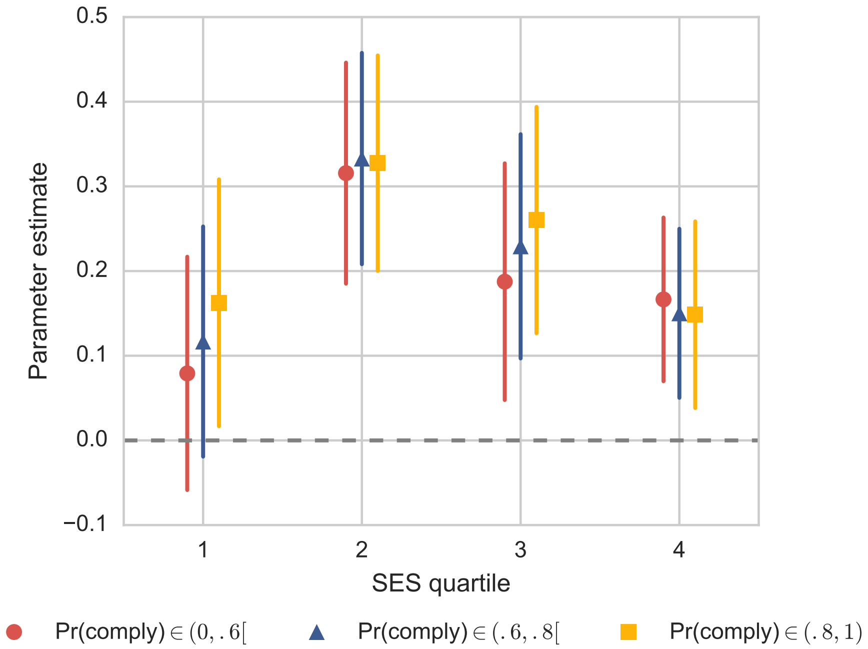


Figure 9: Effects of MSS on actual peer composition, split by expected compliance and household SES

The parameters are estimated in a model with three way interactions between prediction group dummies, SES quartile dummies and MSS as a continuous measure. Two-way interactions between prediction ventiles and SES ventiles and interactions between non-Western and gender dummies are included as controls. The dependent variable is the average SES in the cohort belonging to the school in which the child enrolls.

dren enrolled in primary school.

The results of this estimation are displayed in Figure 9 where the interactions between the mechanical SES, household SES-quartile and prediction quartile are plotted. All interactions but for the combination of low SES and low predicted compliance are positively significant, meaning that the first stage in an IV would have predictive power. In other words, the policy of boundary changes actually affect all subgroups to some degree.

There is heterogeneity across both predicted compliance and household SES. The general pattern along the SES dimension follows the shape of the out-of-sample prediction in Figure 4a.

The highest prediction group within the lowest SES-quartile has a higher estimate than the other groups within the same SES-quartile. This pattern provides supporting evidence that the heterogeneity of our estimated ITT-effects is actually due to exposure to the policy of boundary changes.

For the highest SES quartile, we see relatively smaller first stage coefficients, compared to the middle of the SES distribution. However, these are significant and positive, which imply that the zero effect of MSS cannot readily be explained by non-compliance. This implies that, even though some households may choose to opt out, the policy is actually effective in manipulating the actual peer composition of children, regardless of socioeconomic status. This mirrors our conclusions in Bjerre-Nielsen and Gandil (2018), that overall there is a degree of compliance when attendance boundaries change.

The large “first stages” in Figure 9 combined with the very small reduced-form estimates for the high-SES children in Figure 8 indicate that, even though high-SES groups are exposed to changes in the socioeconomic makeup of their peers, this does not meaningfully affect their performance on the Danish language tests.

The insignificant first stages for the lower prediction groups in the bottom of the SES distribution, however, point to the limits of SAB changes as a policy tool. We here identify a disadvantaged group, which is not affected by a policy, for which they may be the primary target. Though we cannot estimate it, we conjecture that these children may gain from being exposed to stronger peers, but this policy may not effectuate the intended exposure due to low compliance.

7.3 Missing outcomes

As mentioned in section 3, we only obtain test scores on the children enrolled in public schools. In other words, for all children enrolled in private schools we do not observe an outcome. If some children move to the private sector as a response to changes in the expected sociodemographic compo-

	(1)	(2)	(3)	(4)	(5)
MSS	0.142	0.226*	0.213**	0.231**	0.230**
	(0.168)	(0.104)	(0.0711)	(0.0723)	(0.0728)
MSS x SES Q2	-0.214*	-0.132	-0.111†	-0.129*	-0.129*
	(0.0942)	(0.0813)	(0.0608)	(0.0609)	(0.0603)
MSS x SES Q3	-0.349***	-0.130	-0.138*	-0.183**	-0.172**
	(0.0956)	(0.0789)	(0.0587)	(0.0612)	(0.0573)
MSS x SES Q4	-0.330**	-0.160*	-0.207***	-0.208***	-0.247***
	(0.106)	(0.0807)	(0.0576)	(0.0590)	(0.0573)
Outcome	TS missing	TS	TS	TS	TS
Imputation			Predicted TS	Pr(comply)=0	Pr(comply)=1
N	15648	11669	15648	15648	15648

Standard errors in parentheses
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Effect of mechanical SES changes on Danish test scores, imputed outcomes

This table present robustness tests. The dependent variable in column one is a dummy for whether the test score is missing. Column 2 repeats the results of column 2 in Table 2. Column 3-5 present the results when using different imputation techniques. Column 3 use the random forest prediction model to impute unobserved outcomes. Column 4 and 5 also impute using the prediction model but synthetically setting the probability of complying by the school assignment to 0 and 1 respectively.

sition, then our results could be biased. In the present framework, we can illustrate this by regressing an indicator for missing outcome on the independent variables. The result can be read from column 1 in Table 3. We see that sensitivity is increasing in household-SES. For the lowest SES-quartile, we find an insignificant effect of 0.142. However, the sum of the base and interaction parameters become negative for the higher SES groups. In other words, the more positive a change in MSS the lower the probability of a missing test score. There is, therefore, indications of non-random attrition and we need to test whether our results are sensitive to the missing outcomes. We impute the missing test-scores for pupils in a number of ways.

First, we use a similar approach as when we predict baseline compliance. We use a Random Forrest regression on the training sample with the same background variables as for predicting compliance to predict test scores. However, as we never observe test scores for children enrolled in private school we can only train the model on the children enrolled in public school. We include the predicted compliance rate in the model. Thus, the model may choose to let predicted test score be a function of the predicted compliance. We once again test our model against the test dataset and find that the model fits the mean well. We then use the model to predict the missing outcomes in our causal sample and include these in the regression of test scores on mechanical changes in SES.

Doing this, we assume that the children going to private school would do as well as the observably equivalent children in the public school system absent a change in characteristics. For the households who would always choose private school, imputation should matter little. For the "defiers", which are moved to private school because of the change in MSS, we assume that the choice of private school is able to compensate fully for the effect of a change in SAB on the performance of the child.

The results of this exercise are seen in column 3 of Table 3, where we also repeat the baseline estimation in column 2 for comparison. We see almost no change in the results. To further test the sensitivity, we estimate the model two times, but now we artificially change the probability of compliance to zero and one respectively. We then use the predicted test scores with this artificial probability for imputation. The results barely budge. We also estimate a two-step Heckman correction model, where we include the baseline prediction in the first stage. Appendix Table A1 shows a wholly insignificant Mills ratio and estimates completely in line with the baseline results.

Using our three imputed test scores we perform our three-way interactions of the effects corresponding to Figure 7 in Appendix Figure B.2. We find that our conclusions are robust to the imputation. Overall, the robustness results provide no indication that our results should be biased due to non-random attrition. A reason could be, that we are fully able to control for the selection by including covariates. In other words, we do not find evidence that households sort into private school on unobservables.

8 Conclusion

Peer effects can potentially be important for understanding how inequality is transmitted from one generation to the next. In this paper, we provide evidence that children from disadvantaged backgrounds benefit from exposure to children from more advantaged backgrounds. Importantly, privileged children do not seem to be affected at all. This implies that associational redistribution may increase equality and economic efficiency at the same time. However, the analysis also elucidates the limitations of such policies. Compliance is crucial for the intended effect to materialize. In other words, the institutional context wherein authorities seek to manipulate student bodies matter for the expected effects.

As compliance is not random, it also underlines the difficulties going beyond the reduced form when estimating peer effects. A contribution of this paper is therefore also to show the limits to estimating causal peer effects in a context where compliance is not ensured.

References

Angrist, J. D. 2014. The perils of peer effects. Labour Economics, 30, 98–108.

Baum-Snow, N., Lutz, B. F. 2011. School desegregation, school choice, and changes in residential location patterns by race. American Economic Review, 101, 3019–46.

Bjerre-Nielsen, A., Gandil, M. H. 2018. Defying attendance boundary policies and the limits to combating school segregation.

Blume, L. E., Brock, W. A., Durlauf, S. N., Jayaraman, R. 2015. Linear social interactions models. Journal of Political Economy, 123, 444–496.

Bramoullé, Y., Djebbari, H., Fortin, B. 2009. Identification of peer effects through social networks. Journal of econometrics, 150, 41–55.

Breiman, L. 1984. Classification and regression trees. Wadsworth statistics/probability series, Wadsworth International Group.

Breiman, L. 2001. Random forests. Machine learning, 45, 5–32.

Carrell, S. E., Sacerdote, B. I., West, J. E. 2013. From natural variation to optimal policy? the importance of endogenous peer group formation. Econometrica, 81, 855–882.

Chetty, R., Hendren, N., Kline, P., Saez, E. 2014. Where is the land of opportunity? the geography of intergenerational mobility in the united states. The Quarterly Journal of Economics, 129, 1553–1623.

Coleman, J. S. et al. 1966. Equality of educational opportunity..

Durlauf, S. N. 1996a. Associational redistribution: A defense. Politics & Society, 24, 391–410.

Durlauf, S. N. 1996b. A theory of persistent income inequality. Journal of Economic growth, 1, 75–93.

Epple, D., Romano, R. E. 1998. Competition between private and public schools, vouchers, and peer-group effects. American Economic Review, 33–62.

Graham, B. S. 2018. Identifying and estimating neighborhood effects. Journal of Economic Literature, 56, 450–500.

Gruber, J., Mullainathan, S. 2006. Do cigarette taxes make smokers happier? In Happiness and Public Policy, Springer, 109–146.

Hanushek, E. A., Machin, S. J., Woessmann, L. 2016. Handbook of the Economics of Education. Elsevier.

Hoxby, C. M., Weingarth, G. 2005. Taking race out of the equation: School reassignment and the structure of peer effects.Technical report, Working paper.

Imberman, S. A., Kugler, A. D., Sacerdote, B. I. 2012. Katrina's children: Evidence on the structure of peer effects from hurricane evacuees. American Economic Review, 102, 2048–82.

Manski, C. F. 1993. Identification of endogenous social effects: The reflection problem. The review of economic studies, 60, 531–542.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Sacerdote, B. 2011. Peer effects in education: How might they work, how big are they and how much do we know thus far? In Handbook of the Economics of Education, 3, Elsevier, 249–277.

Wager, S., Athey, S. forthcoming. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association.

A Why not IV?

An implication of monotonicity is stochastic dominance of distributions of the endogenous treatment for different values of the instrument. In our context, this would imply that the distribution of values of the actual peer composition for a value of mechanical peer composition either dominates or are dominated by the distribution for another mechanical peer composition. We can check for this in our data. We group values of MSS (after demeaning out the fixed effects) into quartiles and for each quartile plot the empirical cumulative distribution functions (cdf) of cohort compositions. The result of this exercise is displayed in Figure A.1. It is immediately evident that the cdf of the two lowest quartiles cross the other cdfs, which mean that stochastic dominance is not satisfied. This implies that for sufficiently low values of the instrument some children will actually experience *higher* values of peer compositions. This could be the case if parents for a sufficiently low expected peer composition opt out of their designated school into another school with higher peer composition than baseline. We show in a Bjerre-Nielsen and Gandil (2018), that this is indeed the case. From this follows that some weights in the calulated average treatment effect will be negative. The usual result, that the IV measures an average of a heterogeneous effect on compliers, does therefore not hold in our context.

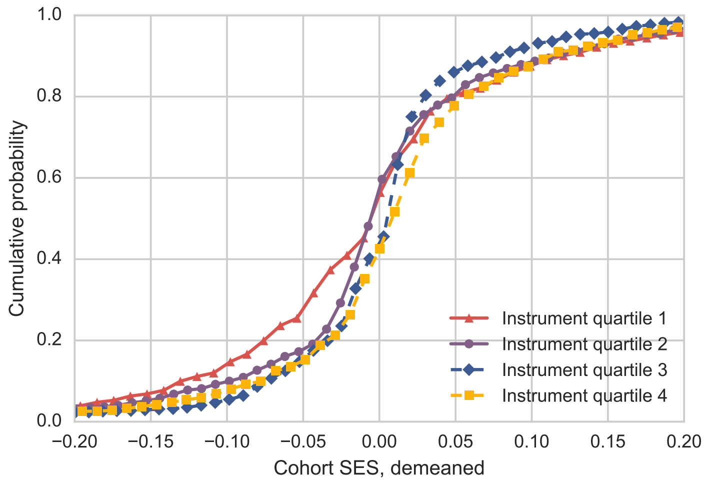


Figure A.1: CDF of treatment for quartiles of instrument variables

The figure present the cumulative distribution functions for demeaned treatment for quartiles of demeaned mechanical SES. A necessary condition for monotonicity to be valid is stochastic dominance. This is clearly not fulfilled as the red density intersects the other densities.

B Additional tables and graphs

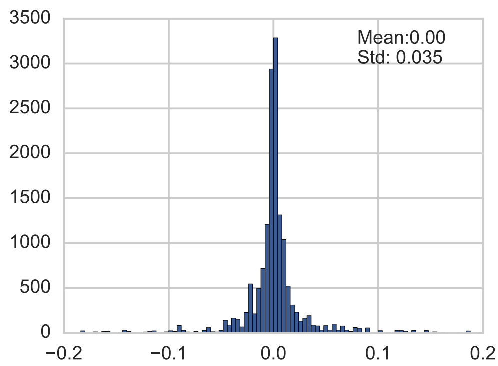


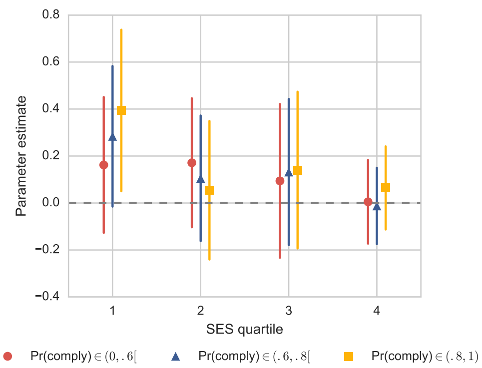
Figure B.1: Demeaned mechanical SES

	Test score	Missing
MSS	0.229*	-0.477
	(0.0900)	(0.440)
MSS x SES Q2	-0.137†	0.877*
	(0.0794)	(0.391)
MSS x SES Q3	-0.137	1.390***
	(0.0841)	(0.388)
MSS x SES Q4	-0.167*	0.893*
	(0.0819)	(0.368)
Mills	-.013	
Mills SE	.07	
Obs.	15648	

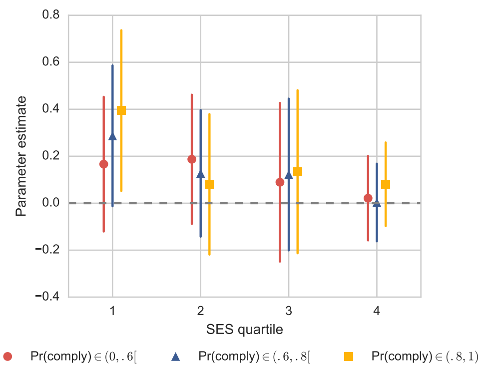
Standard errors in parentheses
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1: Heckman 2-stage sample correction model

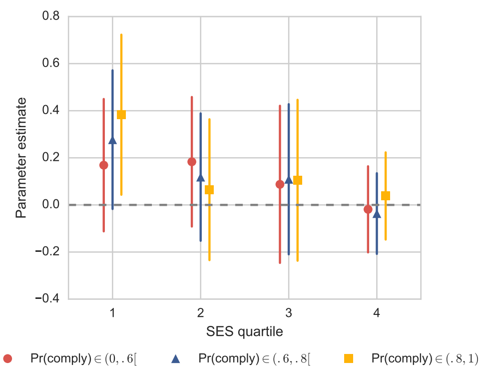
The model is estimated correspondingly to the models in Table 3. SAB-year fixed effects are included in the estimation.



(a) Imputed outcomes



(b) Imputed with P(comply) = 0



(c) Imputed with P(comply) = 0

Figure B.2: Robustness of estimates

The parameters are estimate in a model with three way interactions between prediction group dummies, SES quartile dummies and MSS as a continuous measure. Two-way interactions between prediction vintiles and SES vintiles and SES-SAB-year fixed effects are included as controls. The dependent variable is test scores in Danish language taken in second grade. The imputation method used can be read from the captions of the subfigures.

Appendix A

Technical appendix: Privacy in spatial data

Privacy in spatial data with high resolution and time invariance

Andreas Bjerre-Nielsen & Mikkel Høst Gandil

Abstract

This paper presents a graph-based algorithm to construct partitions of data which preserve k-anonymity while maximizing precision. We present the algorithm and apply it to the geographical population distribution of Denmark. The algorithm provides a time-stable geographic partition with seven times larger precision than the smallest partition available so far. All software is made available free of charge on the authors' websites.

1 Introduction

Researchers working with data on individuals or organizations often work under the constraint of having anonymized entities. The anonymization works by hashing names, social security numbers etc.. One major problem often faced by researchers is not having access to the exact address and home location of individuals in the data.[1] The lack of access to individual location implies that researchers have to resolve to crude measures e.g. administrative boundaries. These measures often suffer from lack of precision and instability over time. This prohibits computation of geographical patterns and meaningful interpretation of the effects of nearby amenities.

Much of the existing literature focuses on making computational systems that can anonymize user location data over time such as contin-

[1] The reason for lack of access to such information is that it can be used to reverse engineer the identity of an individual by combining searches on local address and phone directories and/or social media.

ual GPS data from smartphones. See Gedik & Liu (2004), Mokbel et al. (2006) for examples of such approaches. In this paper, we present an algorithm, which creates a mapping from individual location to an approximate location shared by at least k other individuals. This aim is shared by earlier work on anonymization but our approach differs by considering two additional constraints. The first extra condition is that the mapping of location to approximate location is constant over time. The second is that instead of using exact individual locations as input it takes an approximate location corresponding to squares in a grid.[2]

Our contribution is to provide a procedure that makes spatial clusters that are stable over time and contain a minimum number of inhabitants over time according to a local optimization. The local optimization ensures high precision while respecting privacy. We measure location privacy at the *k-anonymity* level, see Samarati & Sweeney (1998). This implies that the location of one individual is the same as for at least k-1 other individuals.

This paper is structured the following way:

- We provide a method that computes a locally optimized procedure that produces k-anonymous partitions using a pre-existing set of polygons. This method works also for requiring k-anonymity across multiple points in time.
- We apply this method in the context of Danish data in the context of the Danish Squarenet. This data is a division of Denmark into a grid of equally sized squares of a certain length. We focus on the setting where cells measure 100m by 100m which has the cartographical reference “DKN_100m_euref89”.[3] We partition this grid into polygons which are collections of cells such that each collection polygon has at least 100 inhabitants each year it is inhabited.
- We illustrate the results with a handful Danish municipalities.

One advantage of using the Danish Squarenet (Det Danske Kvadratnet) is that it is based on a precise metric coordinate system and is sup-

[2]This condition is necessary for working with Danish registry data as the number of people living at a given address is too sensitive for researchers to access.

[3]http://www.dst.dk/da/TilSalg/produkter/geodata/kvadratnet

ported by various governmental organizations and can be merged unto other geodata sources such as Open Streetmaps. Moreover, the grid is stable over time and exogenous with respect to political and deistic partitions (such as municipalities and parishes) alleviating fears of endogeneity of geographic subdivisions. The choice of 100 k-anonymity for each year was chosen by Statistics Denmark to ensure a modest level of privacy given that researchers also have access to multiple other identifiers. Given this constraint, our approach improves our measure of spatial precision on average from 3700m for parishes to 510m; moreover at the median our measure of spatial precision is reduced to 300m from 3500m.[4]

The main thrust of the problem is that the constraint of a minimum number of inhabitants creates a trade-off between precision and inclusion. Larger polygons that may include more people, and thus make a larger share of the population available for analysis. However, geographical precision is a decreasing function of polygon size. In other words, the goal is to have as *many polygons* with as *few people* in them while upholding the restriction of *at least 100 people*are in each polygon, i.e. group of squares.

This paper documents our approach to this problem. The algorithm is implemented in Python and the associated code is available in this public git repository and distributed under MIT licensing. The algorithm presented here is simplified to clarify the main logic. In practice, a number of steps is taken to simplify the "geography" of the data. We refer to the code repository for these procedures.

This paper proceeds as follows. We begin in Section 2 by describing the general problem and the metrics, we use for diagnosis. We then proceed in Section 3 to describe the algorithm before giving an example of the result from running the algorithm. Section 6 concludes.

2 General approach

In this section, we present the problem and the logic of our algorithm. The basic data consists of a list of square cells. Each square is associated

[4]We measure spatial precision as the square root of the area in the convex hull of the shape - see Section 4.

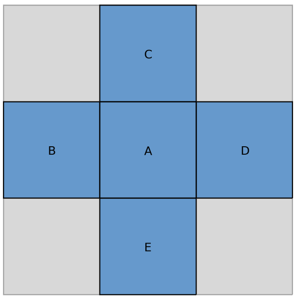


Figure 1: Possible neighbors to A

with a vector where each entry contains the population in a given year. The goal is to connect these squares to fulfill a restriction on minimum of the sum of population vectors. We restrict the possible connection by requiring that two squares must share a border in order to be connected as illustrated in figure 1. Thus each square has four possible connections (if figure 1 A can connect to B,C,D,E).

For a *given year* there are a number of different combinations. For simplicity, we restrict the example to a 3×3-square. An example of different feasible partitions for a given population distribution is illustrated in figure 2. There are many possible combinations and we need a method to rate these partitions against each other. Here we face a trade-off between inclusion of a large share of the population and geographical precision for those individuals included.

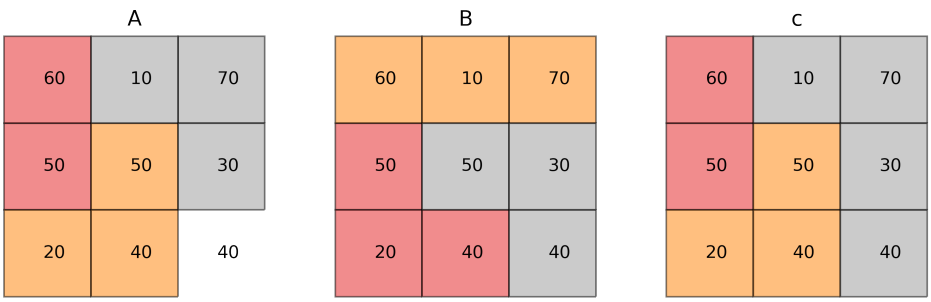


Figure 2: Examples of partitions

To be precise we introduce the following notation. Define the set of squares $\mathcal{S}$, indexed by j. Each square is defined by a coordinate of the lower left corner, (x_j, y_j), and a population n_j.[5] Let partition $\mathcal{P} \in \mathcal{S}$ be a set of non-overlapping polygons, indexed by p containing a number of squares.

As a measure of inclusion, we simply take the share of the population not included. We thus define:

$$non\_pop_{\mathcal{P}} = 1 - \frac{\sum_{p \in \mathcal{P}} \sum_{j \in p} n_j}{\sum_{j \in \mathcal{S}} n_j}, \tag{1}$$

where p is a polygon in partition $\mathcal{P}$, $p \in \mathcal{P}$. The polygon can itself be thought of as a set of geographically connected squares. As a measure of precision in a polygon, we use the bounding box of the polygon and take the length of the diagonal.

$$dist_p = \sqrt{[max_{j \in p}(x_j) - min_{j \in p}(x_j) + 1]^2 + [max_{j \in p}(y_j) - min_{j \in p}(y_j) + 1]^2} \tag{2}$$

When evaluating the precision of a partition $\mathcal{P}$ we calculate the weighted distance, simply expressed by:

$$weighted\_dist_{\mathcal{P}} = \frac{\sum_{p \in \mathcal{P}} \sum_{j \in p} dist_p \times n_j}{\sum_{j \in \mathcal{P}} n_j} \tag{3}$$

To evaluate the trade-off we define a simple linear loss-function:

$$L(\mathcal{P}) = -non\_pop_{\mathcal{P}} - \beta \times weighted\_dist_{\mathcal{P}}, \tag{4}$$

for some weight β. In table 1 it can be seen that with these diagnostics one should prefer partition B. One can see that partition C does cover the entire population it is punished by the relatively large distance in the gray group compared to partition B where the polygons are “more convex”.

[5] In this example we only have one year of analysis. If we need to consider the k-anonymity restriction in multiple years n_j would be a vector.

P	dist	weighted_dist	pop_weight	loss
A	7.89	2.35	0.89	-0.1316
B	8.82	2.95	1.00	-0.0295
C	8.67	2.97	1.00	-0.0297

Table 1: Example of diagnostics of partitions in figure 2, $\beta = 0.99$

3 Method

In this section, we go into more detail about how data has been processed and how the algorithm works. For further details see the Python code available in our git repository.[6] The assignment of spatial cells to a collection consists of three overall stages:

1. Pre-processing the data
2. Application of the algorithm
3. Post-processing by merging the best partition for sub-areas into one large coherent partition

We move to briefly describe the stages and put an emphasis on the main algorithm.

3.1 Preprocessing

The first stage consists of processing data which is repeatedly used in the computation. A first stage is to split Denmark into sub-areas. Application of the algorithm below is time-consuming and it runs slowly, especially on larger areas. In order to speed up the process we to divide Denmark into sub-areas. We employ the (current) Danish municipalities as our basis for this division. An issue with using municipality data is that there is a big difference between the smallest municipality (Frederiksberg) and some of the larger rural ones. Therefore we begin our analysis by splitting the large municipalities (>25 square kilometers) into smaller chunks. This is done by expressing the square net cells as

[6] `https://github.com/abjer/privacy_spatial`. For the practical implementation of our code we have relied on various open source projects, especially Blondel et al. (2008), Csardi & Nepusz (2006), McKinney (2011), Schult & Swart (2008), Van Der Walt et al. (2011)

a network graph where links/edges exist between cells within a certain distance. This representation allows us to use the Louvain method, see Blondel et al. (2008), to make a pre-division of larger municipalities into smaller and manageable chunks. In other words, we construct connected components of cells and run the algorithm within these components.

The next pre-processing stage concerns identifying newly constructed dwellings. For the requirement by Statistics Denmark, to have at least 100 people in a populated polygon, to be fulfilled every year adds complexity. When cells in a polygon exhibit large variation in the population, this can become a problem. Especially for new construction or data-breaks where each cell can go from having no population to a possibly large population. When coupling a group of cells that contain both cells with new construction and cells without it becomes difficult to satisfy the population constraint in the initial years while maintaining high precision. To overcome this, we identify for each cell if there is a break of going from no population to some population. We then apply our method below to cells with such breaks separately. This greatly improves our precision in the resulting partition of the country.

3.2 Basic algorithm

Naturally, the possible valid partitions explode with the number of cells to consider and it is not feasible to check them all to achieve optimality. Instead, the basis algorithm runs from a random starting point. Prior to the run of the algorithm, a network graph is constructed wherein feasible neighbor cells are represented by edges.

Every population vector associated with each cell, $\mathbf{n}_i$, has the same length and may contain zeros. Thus on the outset, we have the following data available:

- $\mathcal{S}$: Set of squares, where each square, i is defined by the tupple $(x_i, y_i, \mathbf{n}_i)$.
- $\mathcal{G}$: A graph with information on which squares are neighbors.

Furthermore, we define a number of variables:

- R: an integer, the maximum attempts an algorithm should perform before exiting.
- r: number of attempts, initially set to 0.
- $\mathcal{U}$: A set of unassigned squares, initially equal to $\mathcal{S}$.

The basic algorithm is displayed in Algorithm 1. When all squares are assigned or the sum of the remaining squares is less than 100 the algorithm stops.

3.3 Application of algorithm

We apply our algorithm for each of the sub-areas described above. Each application of the algorithm consists of multiple stages. The first step is to apply the algorithm to year breaks caused by construction (see preprocessing above). If there are more than enough inhabitants collectively, these are added to the partition. The second is to apply the algorithm to all the squares of the sub-area, which are not assigned in the first stage (with year breaks). Finally, we apply an additional algorithm, that check for 1:1 feasible exchanges of squares between neighboring polygons. The condition for an exchange is that it improves the distance between the polygon cells without violating the feasibility constraints.

Upon termination, we compute various measures for the partition and these are stored together with the partition. The algorithm runs multiple times for each sub-area and we pick the partition with the numerically smallest value of our loss function defined in (4). The definition of $\mathcal{S}$ is important for the amount of time the algorithm takes to finish. Densely populated areas are partitioned quickly, but in geographically large and sparsely populated areas, the algorithm will take a very long time to complete as the algorithm does not scale well. Therefore, it is essential to split the spatial areas where the algorithm is executed into smaller parts.

3.4 Finalizing the output

After having executed the algorithm multiple times on all the sub-areas the output is collected. Among the output, the candidate sub-partition

Data: $\mathcal{S}, \mathcal{G}, \mathcal{U}, r, R, \mathcal{P} = \emptyset$
Result: $\mathcal{P}$

```
while (r < R) & (min_A (∑_{l∈U} n_l) ≥ 100) do
    Pick random square from U and define p = {i};
    if min n_i ≥ 100 then
        P ← P ∪ p;
        U ← U\p
    else
        Define F_p as components in G connected to the squares
          already in p;
        if F_p = ∅ then
            r ← r + 1;
            Return to start
        else
            while (min_A (∑_{l∈p} n_l) < 100) & (r < R) do
                Pick random neighbor j from F_p;
                p ← p ∪ {j};
                Update F_p;
                if F_p = ∅ then
                    if min_A (∑_{l∈p} n_l) ≥ 100 then
                        P ← P ∪ p;
                        U ← U\p;
                    else
                        r ← r + 1;
                    end
                    Return to start
                end
            end
            P ← P ∪ p;
            U ← U\p
        end
    end
end
```

Algorithm 1: Basic algorithm

with the lowest value of the loss function trading off spatial precision versus missing data is chosen. Finally, the sub-partitions are merged into one partition.

4 Algorithm output

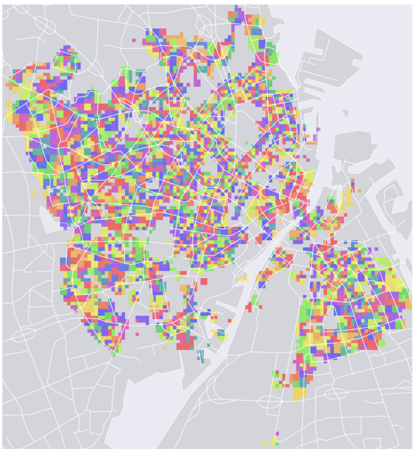
In figure 3 we apply the algorithm to different Danish municipalities. Coloring a map is no simple feature and some polygons will share color while not being joined. Thus, the partition into sub-areas will tend to look slightly worse than it really is.

The strength of the algorithm is evident for densely populated areas as seen in the urban areas in Figure 3a and 3b . It is also evident that in a suburban municipality such as Rudersdal in Figure 3c the algorithm performs well in dense areas, but it branches out once the population density falls and becomes more rural as evident in the North. The branching out, however, becomes more severe as the density falls as seen for the municipality of Vordingborg in Figure 3d. Fortunately, the branching-out covers a relatively small amount of people, and can thus be discarded in the analysis if more precision is called for.

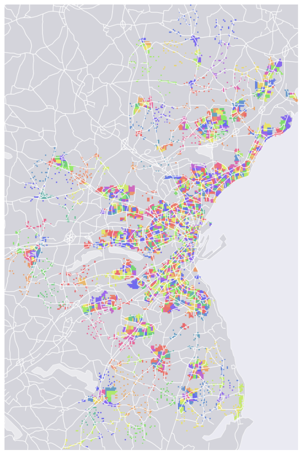
4.1 Descriptive statistics of the partition

Descriptive statistics of a partition of the whole of Denmark is provided in table 2 and Figure 4. The final partition produces more than 30.000 polygons, with a mean of 13 squares in each. It is, however, a right-skewed distribution and the median is thus only 8, as seen in Figure 4. Surprisingly the distribution of areas is bimodal as there is a peak at 1 cell and a peak around 45 cells. This is most likely due to a difference in the structure of large cities and smaller towns.

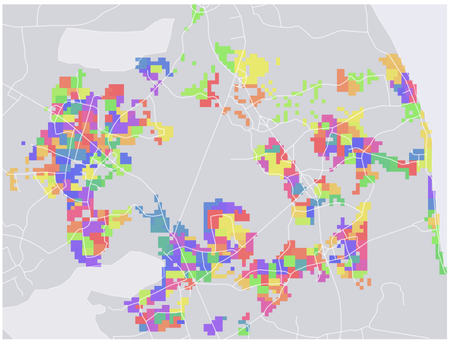
Looking at the scatter-plot between population and area, the scatter plot illustrates the result of the choice of cost function, as there is a trade-off between including the entire population and maintaining precision. The second peak in the distribution of area-size is due to polygons with very low populations. This indicates that the algorithm works as intended.



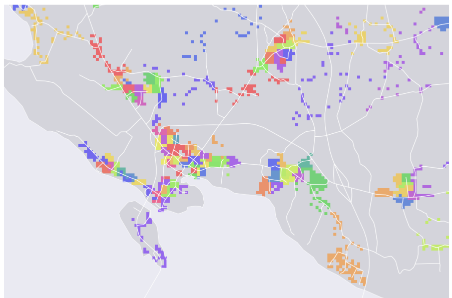
(a) Municipality of Copenhagen



(b) Municipality of Aarhus



(c) Municipality of Rudersdal



(d) Town of Vordingborg

Figure 3: Examples of spatial partitions of suburban and rural municipalities

We can compare the output of our algorithm with the current standard practice for working with geographic data, namely to use the administrative boundaries for parishes. These areas divide Denmark into local districts for the Danish National Church (Folkekirken). The comparison is found in Figure 5. Our measure of distance across the areas shows a seven-fold reduction at the mean and eleven-fold at the median when comparing the population distribution of Danish Squarenet partition to parishes.[7]

As demonstrated above, the size of the polygons is a function of the

[7]The numbers reflect the square root of the area for convex hull of the shape and this proxy for distance; thus for spatial area the improvement are thus respectively 50-fold at the mean and >100-fold at the median.

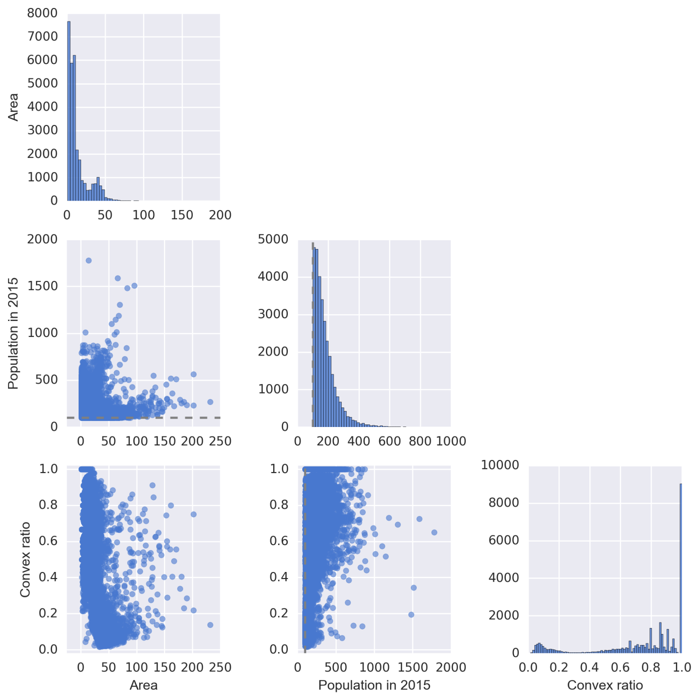


Figure 4: Descriptives of partition

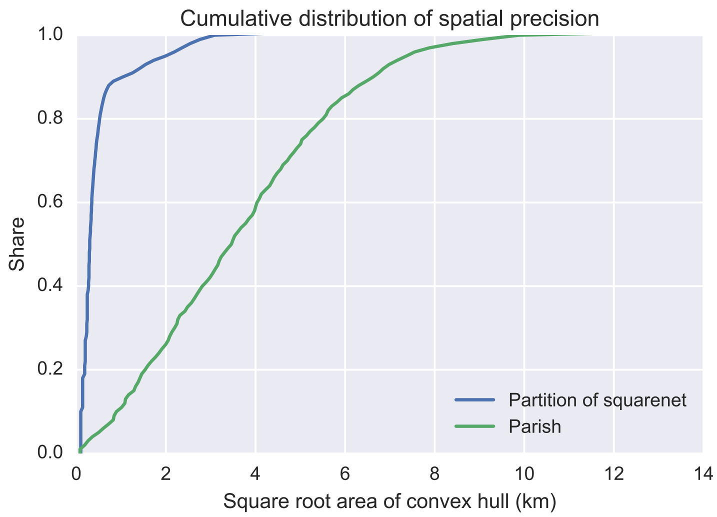
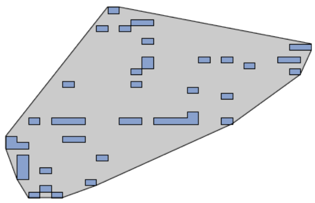


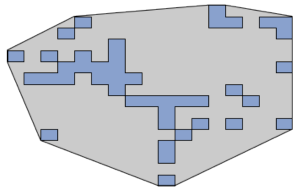
Figure 5: Comparison of spatial precision: square partition vs. Danish parishes

	count	mean	std	min	25%	50%	75%	max
area	30,604	13.84	15.35	1.00	4.00	8.00	17	344
density	30,604	43.01	67.12	0.94	10.56	20.00	39	789
population	30,604	180.85	82.46	100.00	127.00	159.00	208	1,778
convex_share	30,604	0.73	0.31	0.02	0.61	0.84	1	1

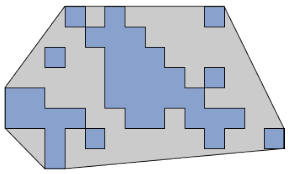
Table 2: Descriptive statistic of full partition



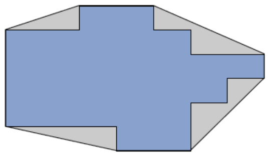
(a) Area/Convex=0.1



(b) Area/Convex=0.2



(c) Area/Convex=0.4



(d) Area/Convex=0.8

Figure 6: Examples of ratios of area to convex hull area

population density. When performing spatial analysis it will sometimes be beneficial to further exclude the polygons, which are very large, as they diminish precision in a given measure. It turns out, that these are identified by having a very low ratio of polygon area to the area of the convex hull. The convex hull is the smallest convex set that contains the polygon. Figure 6 shows an example. A value of around 0.4 seems appropriate for excluding polygons without villages.[8]

5 Example of usage

To briefly illustrate the potential of the algorithm we briefly present a map constructed from register data merged on our final partition of the square-net. Within each polygon, the share employed is registered as well as

[8]This is sometimes referred to as “bizareness”

total population. To maintain a degree of anonymity only polygons with zero *or* at least ten individuals are used. Synthetic individuals are then generated and within each polygon given employment status according to the mean employment rate and a random location within the polygon. The result can be seen in Figure 7 where we use Datashader (Bednar et al. 2016) to aggregate the points to visible pixels. The spatial patterns are very visible. The Northern coastlines of Zealand (right-most island), for example, are dominated by cyan dots representing non-employment. We conjecture that this is mostly pensioners living in a summerhouse whereas bigger cities see a mix of both employed and unemployed. The coastal patterns would not be nearly as visible using other geographic entities, such as parishes, which do not delineate between coast and inland with the same degree of precision.

We note that our approach is limited by the fact that although the partitions are invariant people may relocate and some areas may then lose population. This could imply that some of the sub-areas violate the constraint of having at least k people living in them (100 in the context of Denmark) in the future. We also note that partitioning comes at the cost of a small number of households (<2 pct.) not being assigned an area. These households were mainly rural and they would still have a parish available so a more coarse measure of location would be available.

Figure 7: Employment in Denmark

Red dots represent employed individuals. Non-employed individuals are colored cyan. These individuals may be pensioners and students as well as unemployed. The individuals receive a random location within the polygon of residence. Data is aggregated using the Datashader-package.

6 Conclusion

We have introduced a method that produces a partition of addresses into sub-areas where people reside. The method works by specifying a desired level of privacy to bind while optimizing spatial precision. A key feature of our approach is that the shapes of the sub-areas are time-invariant and thus constant despite changes to administrative boundaries.

We have used our approach to make a partition of the Danish Squarenet. The resulting partition preserves the chosen level of privacy and simultaneously provide a much more accurate spatial precision measured with distance when compared with the de-facto standard measure of using administrative boundaries for parishes. We believe that our approach can be leveraged for similar endeavors in other contexts, e.g. applying it to other countries' addresses in research projects or in business analysis where preserving privacy is required.

References

Bednar, J. A., Crist, J., Cottam, J. & Wang, P. (2016), 'Datashader: Revealing the structure of genuinely big data', *15th Python in Science Conference (SciPy 2016)* .

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008), 'Fast unfolding of communities in large networks', *Journal of statistical mechanics: theory and experiment* (10), P10008.

Csardi, G. & Nepusz, T. (2006), 'The igraph software package for complex network research', *InterJournal-Complex Systems* p. 1695.

Gedik, B. & Liu, L. (2004), A customizable k-anonymity model for protecting location privacy, Technical report, Georgia Institute of Technology.

McKinney, W. (2011), 'pandas: a foundational python library for data analysis and statistics', *Python for High Performance and Scientific Computing* pp. 1–9.

Mokbel, M. F., Chow, C.-Y. & Aref, W. G. (2006), The new casper: Query processing for location services without compromising privacy, *in* 'Proceedings of the 32nd international conference on Very large data bases', VLDB Endowment, pp. 763–774.

Samarati, P. & Sweeney, L. (1998), Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical report, Technical report, SRI International.

Schult, D. A. & Swart, P. (2008), Exploring network structure, dynamics, and function using networkx, *in* 'Proceedings of the 7th Python in Science Conferences (SciPy 2008)', Vol. 2008, pp. 11–16.

Van Der Walt, S., Colbert, S. C. & Varoquaux, G. (2011), 'The numpy array: a structure for efficient numerical computation', *Computing in Science & Engineering* **13**(2), 22–30.