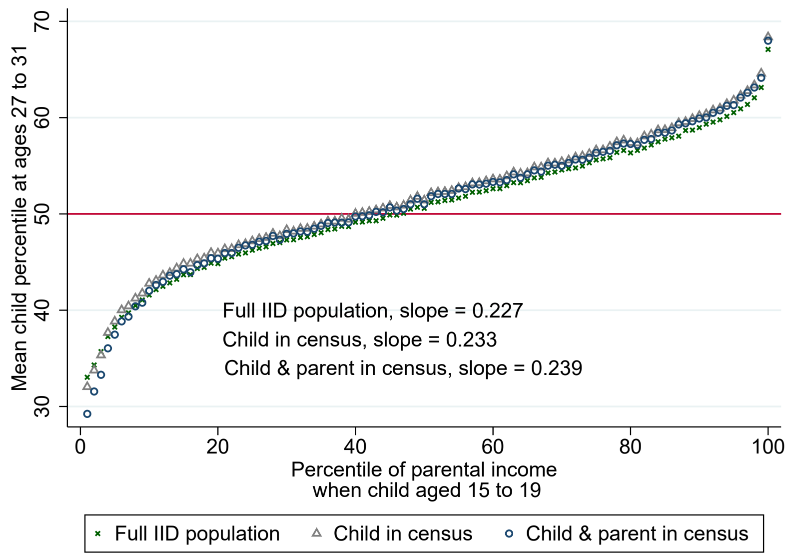
Online Appendix

"Careers and Intergenerational Income Mobility"

Catherine Haeck and Jean-William Laliberte

Appendix Figures

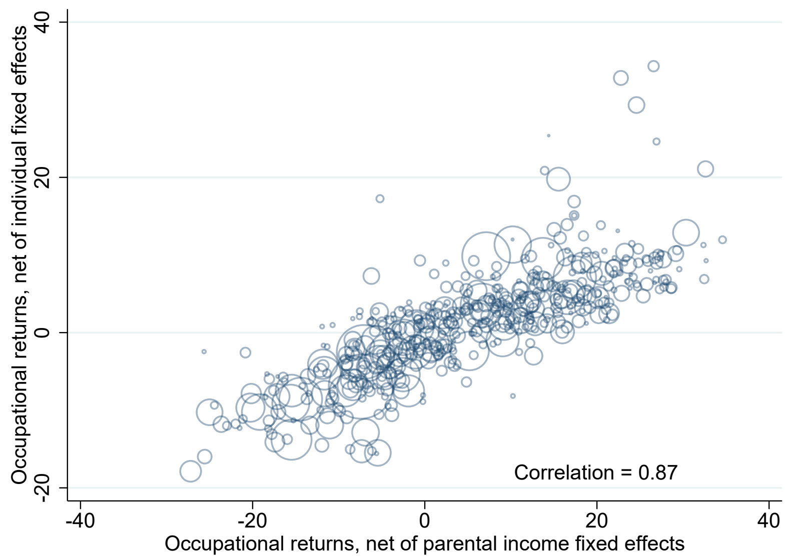
Figure A1: Intergenerational Income Mobility, Sensitivity to Sample Selection



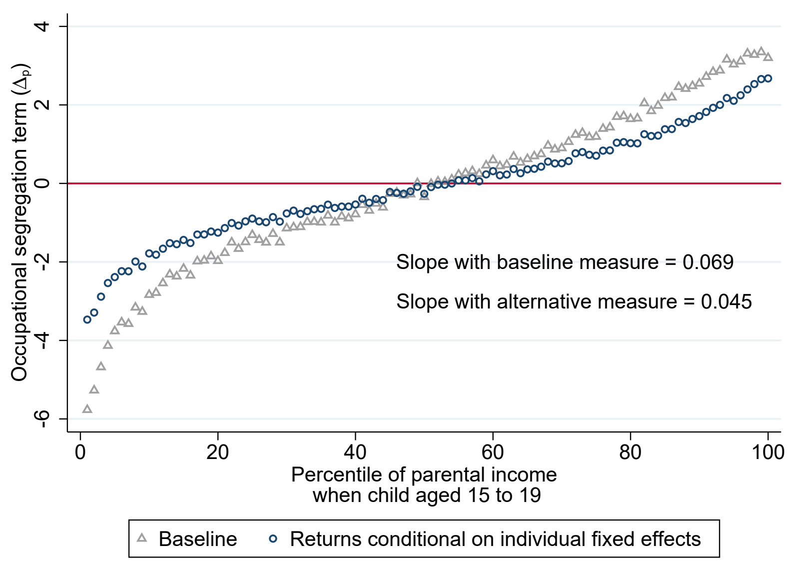
Notes: This figures shows mean child percentile ranks by parental income ranks for three different samples. The green x's show estimates for the entire IID sample. The grey triangles restrict the sample to children who are matched to at least one wave of the Census. The blue circles further restrict the sample to children who have at least one parent matched to at least one wave of the Census.

Figure A2: Robustness to Measurement of Occupational Returns

Panel A: Correlation between baseline and alternative occupational returns

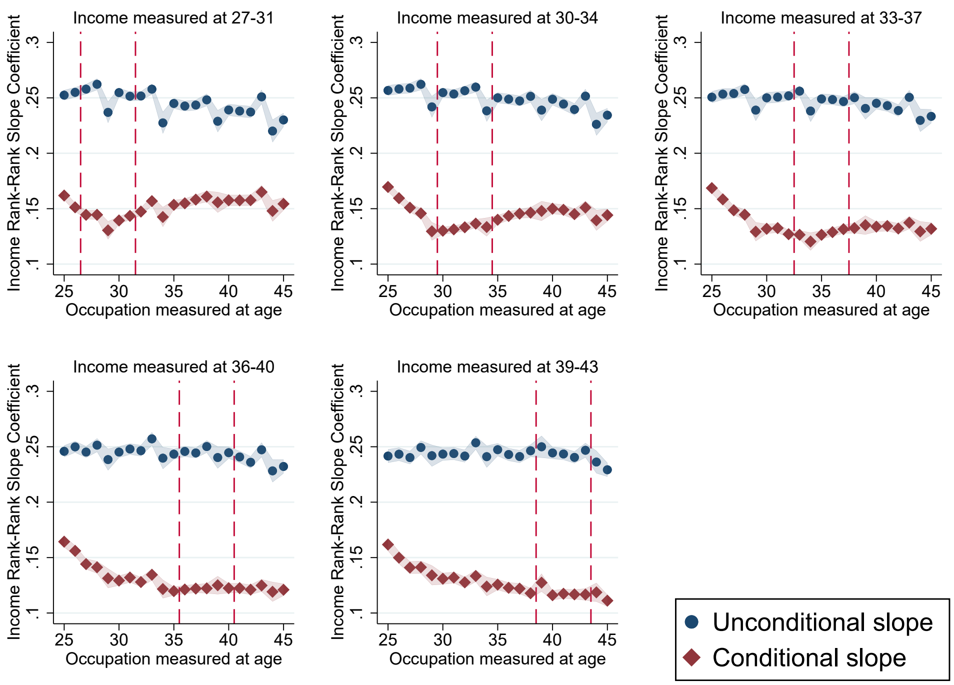


Panel B: Imputed occupational segregation terms for different measures of occupational returns



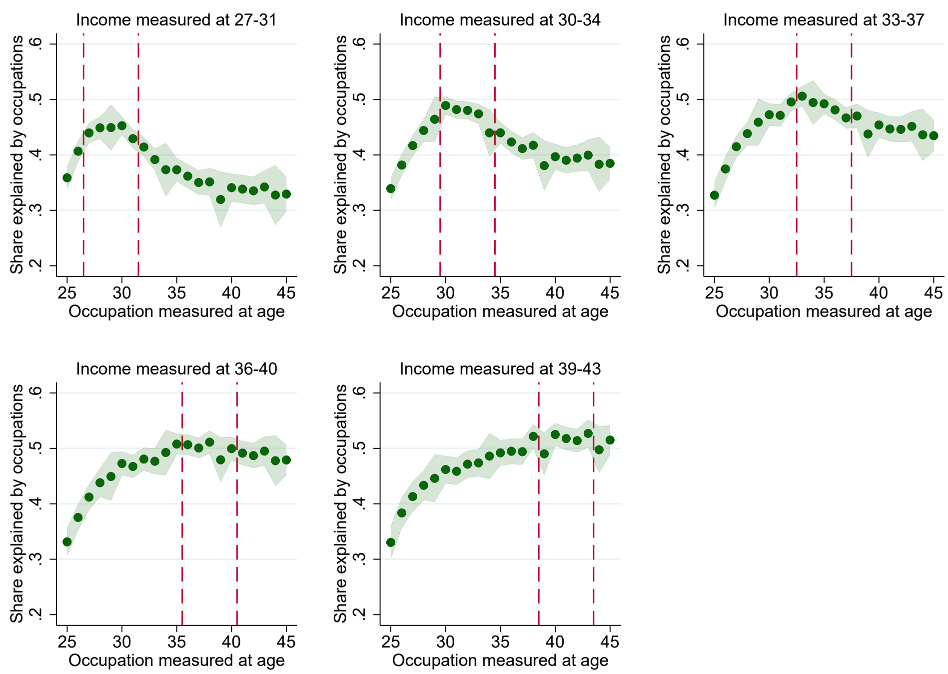
Notes: Panel A plots baseline estimates of occupational returns against alternative estimates based on occupational switches. Each circle represents one occupation, and the size of circles is proportional to the number of children in each occupation. The category "no occupation", which represents about 6% of the sample, is omitted from the scatter plot for visual clarity. Panel B shows values of the occupational segregation term $\Delta_p = \sum_o \hat{\delta}_o s_{o|p}$ for two different measures of occupational returns $\hat{\delta}_o$.

Figure A3: Income Rank-Rank Slopes, Sensitivity to Age-at-measurement



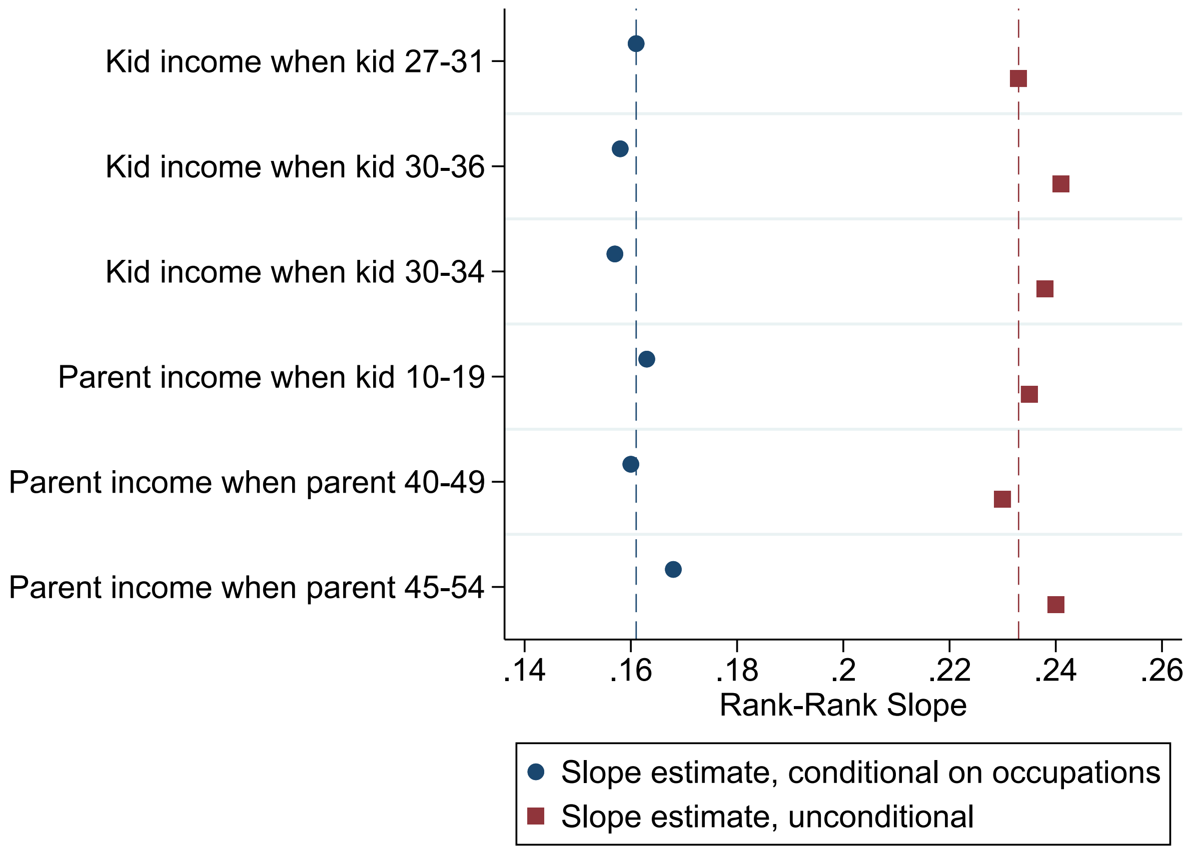
Notes: This figure reports slope estimates of the relationship between child income rank and parental income rank. In each regression, occupations are measured at a specific age (on the x-axis), and income is measured at different age intervals (one panel per age interval). The shaded areas show 95% confidence intervals. The blue dots show unconditional relationships, whereas red diamonds show relationships conditional on a full set of occupation fixed effects.

Figure A4: Share of Rank-Rank Slope Explained by Occupations, Sensitivity to Age-at-measurement



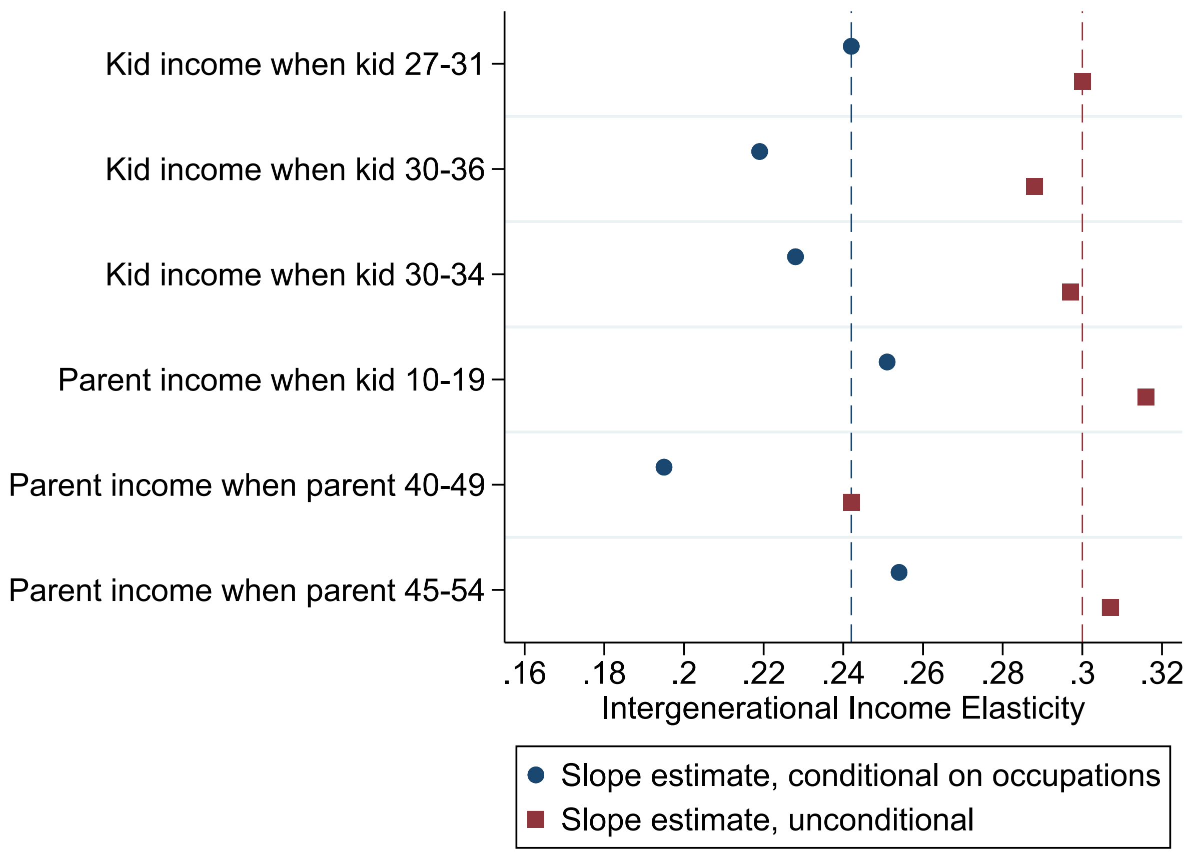
Notes: This figure reports values of $(\beta_U - \beta_C)/\beta_U$, where β_U is the slope of the unconditional relationship between child income rank and parental income rank, and β_C is the slope of the corresponding relationship conditional on a full set of occupation fixed effects. The shaded areas show 95% confidence intervals.

Figure A5: Intergenerational Income Mobility, Robustness



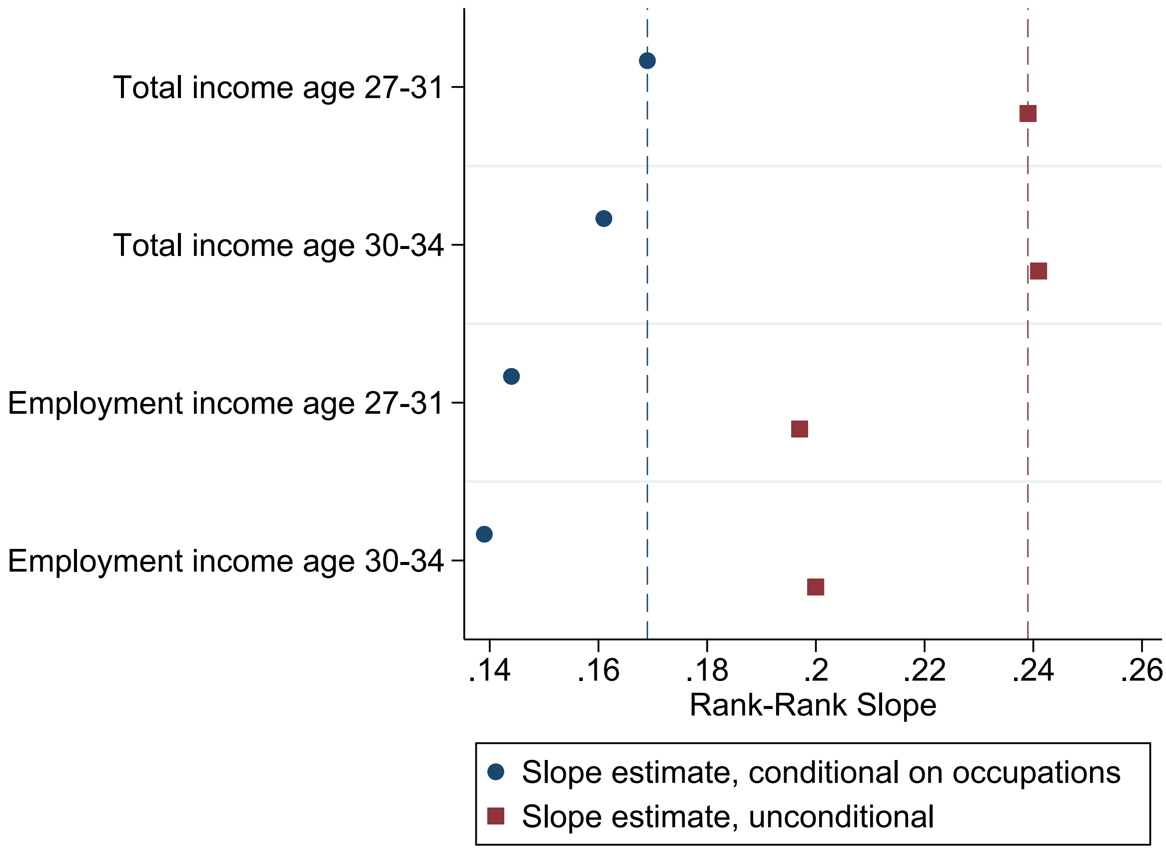
Notes: This figure shows linear slope estimates of the unconditional income rank-rank relationship, and the corresponding relationship conditional on children's occupations. In all specifications the sample consists of all children matched to at least one wave of the Census (i.e., this sample is larger than the main analytical sample that requires parents to also be matched to at least one census). Across rows, we consider different measures of children and parental income ranks. The vertical lines indicate our baseline estimates.

Figure A6: Intergenerational Income Elasticity, Robustness



Notes: This figure shows estimates of the unconditional intergenerational income elasticity, and the corresponding relationship conditional on children's occupations. Across rows, we consider different measures of children and parental income (in log). The vertical lines indicate our estimates for the baseline specification.

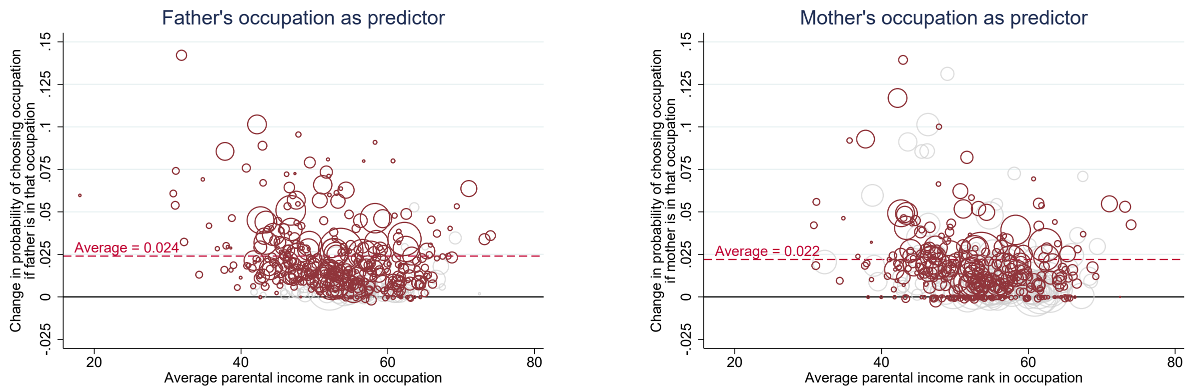
Figure A7: Intergenerational Income Mobility, Children Employment Income



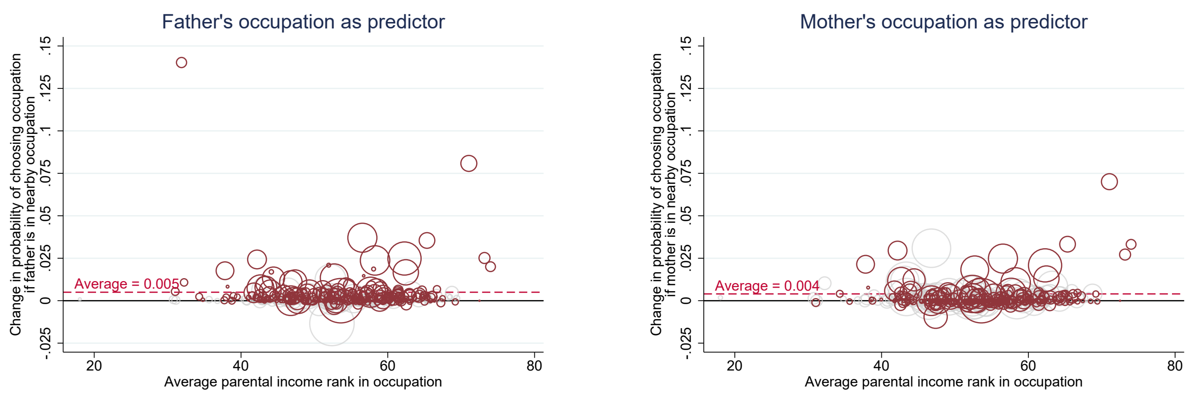
Notes: This figure shows linear slope estimates of the unconditional income rank-rank relationship, and the corresponding relationship conditional on children's occupations. Across rows, we consider different measures of children income ranks. Employment income corresponds to T4 earnings. The vertical lines indicate our baseline estimates.

Figure A8: Occupational Following Across Generations, Separately for Paternal and Maternal Occupations

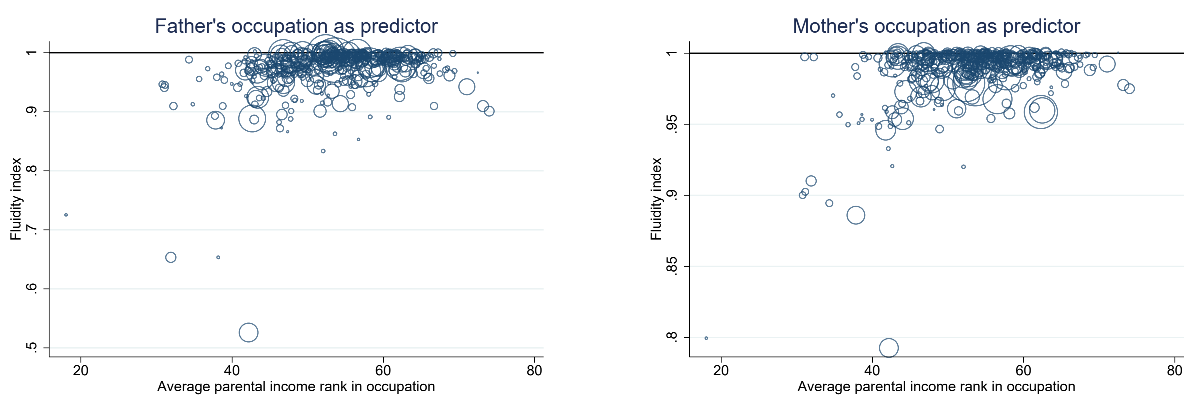
Panel A: Change in probability of pursuing an occupation if parent is in that occupation



Panel B: Change in probability of pursuing an occupation if parent is in similar (but not same) occupation

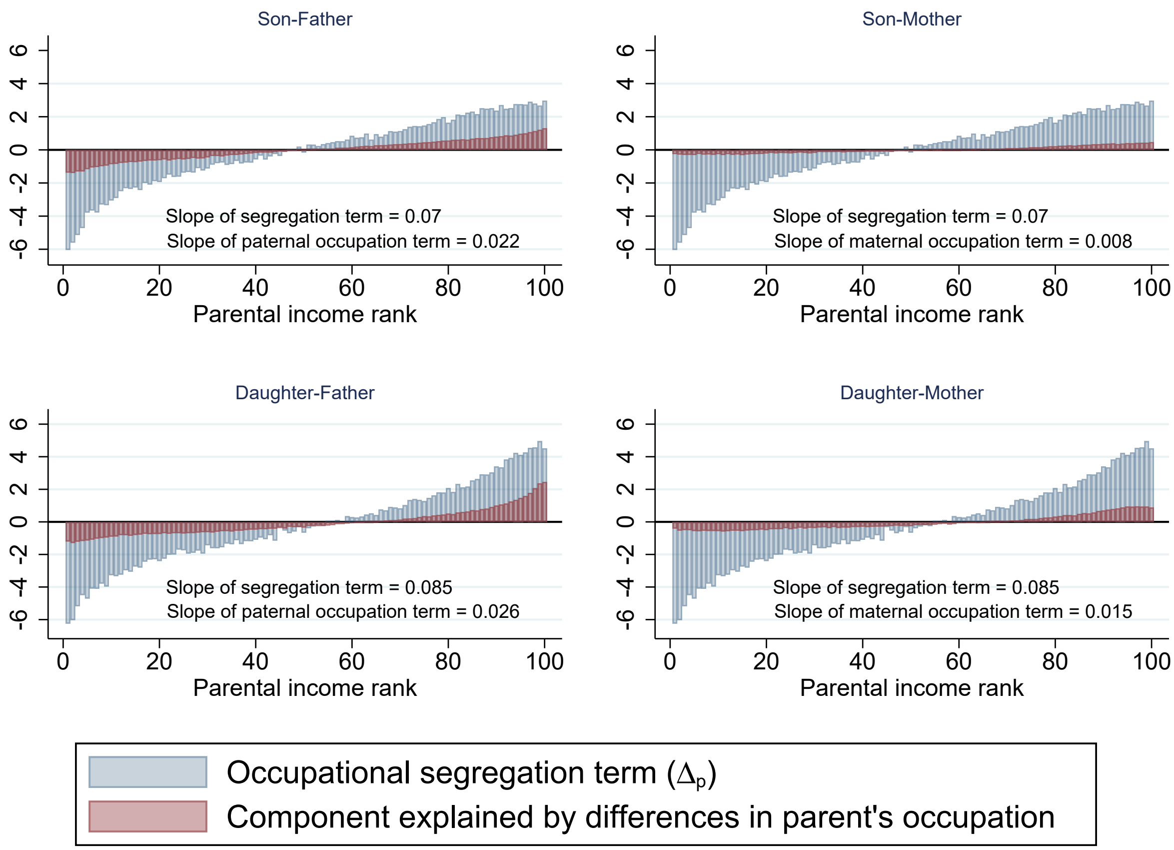


Panel C: Fluidity index by occupation



Notes: This figure reproduces Figure 4, using the father's and the mother's occupation as separate predictors.

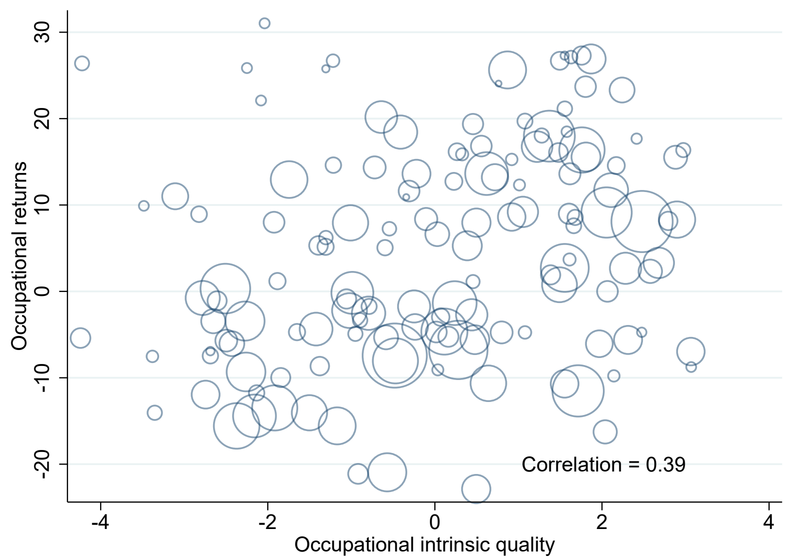
Figure A9: Intergenerational Income Mobility and the Transmission of Occupations, by Gender



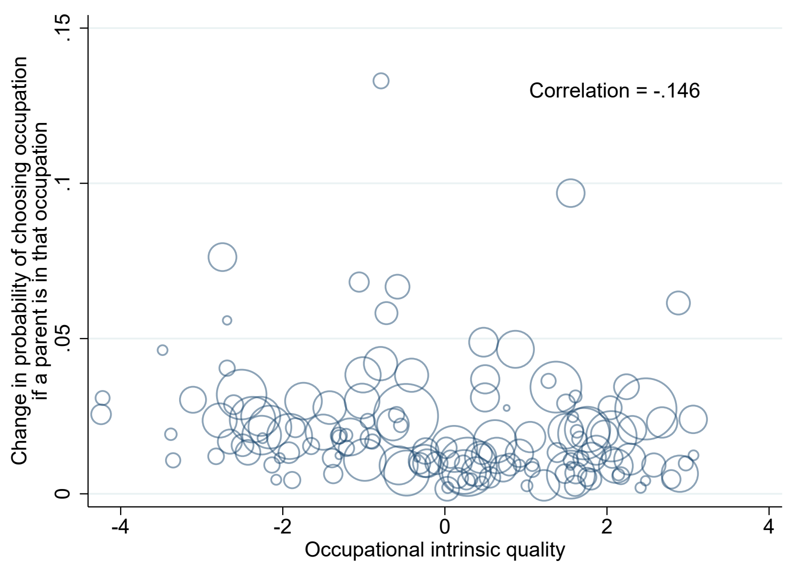
Notes: This figure replicates Figure 4, but conducts the analysis separately by child gender.

Figure A10: Correlation between Occupational Intrinsic Quality and Occupational Characteristics

Panel A: Occupational Intrinsic Quality and Average Occupational Returns



Panel B: Occupational Intrinsic Quality and Occupational Following



Notes: This figure plots intrinsic occupational quality against average occupational returns in panel A, and against estimated coefficients of occupational following (estimates of π_1^o from eq. (3)) in panel B. All variables are collapsed at the 3-digit occupational categories, taking weighted averages across 4-digit occupations within broader categories. The quality variable is based on data from the General Social Survey and methods developed by Boar and Lashkari (2022). Details of its construction are provided in the Data Appendix.

A Data Appendix

A.1 Administrative Data

Sample Selection The analyses are based on a linkage between the Intergenerational Income Database (IID) and six waves of the Canadian Census. We first describe the IID and then how this database was linked with Census data.

The IID is a database constructed by Statistics Canada that links the administrative tax records of children and their parents. This database includes children who were between the age of 16 to 19 in fiscal years 1982, 1984, 1991, 1996 and 2001. The IID therefore covers children born over a 23 year period, from 1963 to 1985 inclusively, but kids born in 1971, 1976 and 1981 are not included. This exclusion was caused by the sampling design of the second wave of the IID (covering birth cohorts post-1970) which had limited funds to produce child-parent links for additional cohorts. Table A1 summarizes the IID cohort structure. It also include population counts and compares them with Census count.

Table A1: The Intergenerational Income Database cohort structure

Match Year	Birth cohorts	Census	IID Count	Ratio	IID Weighted	Ratio
1982	1963 to 1966	1,723,720	1,183,614	0.687	1,517,127	0.880
1984	1965 to 1968	1,563,105	1,124,849	0.720	1,517,126	0.971
1986	1967 to 1970	1,520,745	1,155,248	0.760	1,517,127	0.998
1991	1972 to 1975	1,495,750	1,102,855	0.737	1,484,566	0.993
1996	1977 to 1980	1,570,605	1,166,879	0.743	1,558,393	0.992
2001	1982 to 1985	1,642,535	1,350,222	0.822	1,634,646	0.995

In total, there are over 7 million child-parent pairs in the IID, 5.99 million of which are unique (some birth cohorts overlap across match years, leading to duplicates entries). In the current paper, duplicate entries are excluded. In the IID, parents and children are matched together through the T1 Family File (T1FF) when children are aged 16 to 19 years old. The tax records come from the Canada Revenue Agency and includes all information contained on the tax file. The linkage process is described in details in Corak and Heisz (1999). The match rate, as measured by the IID count divided by the population count of the relevant population in the Censuses, is fairly high, ranging from 68 to 82 percent. Once weights are accounted for, the match rate of the IID count to Census count varies between 88 to 99.5 percent.

The IID records have also been linked with data from the long-form Census using a probabilistic linkage process. The long-form Census covers 1 in 5 Canadian households. It contains information about the demographic, social and economic situation of Canadians. The link has been attempted for both parents and children with Censuses 1991, 1996, 2001, 2006, 2011 and 2016. Each parent or child can be matched with more than one Census if they completed the long-form in multiple years. Depending on the birth cohort, the match rate ranges from 45-58% for the parents (average 53%), with more recent birth cohorts having a lower match rate. For the children, the match rate ranges from 42-62% (average 54%), with more recent birth cohort having a better match rate. The overall match rate for children conditional on also matching at least one of their parents is 31%. This constitutes our main analytical sample.

Children from older cohorts (those born in the early 1960s) have a higher probability of being matched in more than one Census. Using only periods were kids are observed at more or less the same age in the Census would have been too restrictive to allow for a rich documentation of occupational transmission patterns across over 500 occupations. Table A2 highlights in grey the Census waves that were used to retrieve the kids' occupation for each birth cohorts. The average age of children at the time they filled out the census is 32.5 years.

Table A2: Kids occupation and census year by birth cohorts

	Census Year					
	1991	1996	2001	2006	2011	2016
Birth years: 1963 to 1966						
Birth years: 1967 to 1970						
Birth years: 1972 to 1975						
Birth years: 1977 to 1980						
Birth years: 1982 to 1985						

For the parents, we attempted the match for each of the six Censuses and used the most recent Census in which the parent reported an occupation in order to get the most "senior" occupation. We were able to match 53% of the parents to at least one Census. On average, mothers are 57 years old when we link them to Census microdata, and fathers are 59.5 years old. Figures A1 and A5 show that the matched IID-Census subsamples produce patterns of intergenerational mobility that are very similar to those obtained from the entire IID population.

Variables Definitions Occupational codes were harmonized across years using Statistics Canada's concordance tables[27]. Occupation is identified through question 38 in the 2016 questionnaire: "What was this person's work or occupation?". It is then classified using the National Occupational Classification (NOC) or the Standard Occupational Classification (SOC) depending on the Census year. The original data for Censuses 1991 to 2006 is based on the NOC 1991, and was then converted to NOC 2011 codes. Data for Censuses 2011 and 2016 is originally based on NOC 2011 codes. We have a total of 500 harmonized occupational codes.

The main field of study variable comes from question 27 in the 2016 questionnaire. The field of study refers to the predominant discipline or area of study of the respondent's highest completed post-secondary certificate, diploma or degree. This variable was created by assigning a field of study code from the Classification of Instructional Programs (CIP). This question is only answered by respondents who obtained some post-secondary certificate, diploma or degree. For those who did not obtained a post-secondary certificate, diploma or degree, we assigned a "no diploma" category (34% of children in our sample). Over the years, the categories have been slightly modified. We have harmonized them and provide the correspondence table on our website. The CIP classification can be divided into 2-digit, 4 digit and 6-digit codes. We use the 4-digit codes, which results into 446 mutually exclusive categories, including the "no diploma" category.

Parental educational attainment is measured by the highest certificate, diploma or degree. The number and classification of educational categories varies slightly across census waves. To make sure we fully utilize all of the information available in each census, we use a full set of education dummies interacted with census wave dummies to measure educational attainment. We also extract information on place of birth, which lists the country of birth for individuals born outside of Canada, and the province of birth for those born in Canada. The classification of countries of birth varies slightly across census waves, hence we use a full set of place of birth dummies interacted with census wave dummies in our analyses.

A.2 US General Social Survey

The measure of occupational "intrinsic quality" we use in our analyses is based on data from the US General Social Survey (GSS), and methods developed by Boar and Lashkari (2022). The Quality of Work-life Module of the GSS asks respondents about several aspects of their jobs. The index proposed by Boar and Lashkari (2022) combines information from these

[27]https://www.statcan.gc.ca/en/concepts/concordances-classifications

seven questions:

1. At the place where I work, I am treated with respect
2. Does your job regularly require you to perform repetitive or forceful hand movements or involved awkward postures?
3. Does your job require you to do repeated lifting, pushing, pulling or bending?
4. My job requires that I keep learning new things
5. I have an opportunity to develop my own special abilities
6. I get to do a number of different things on my job
7. My job requires that I work very fast.

The estimation procedure is as follows: for each of the seven characteristics, respondents' answers are regressed on occupation fixed effects, and a vector of control variables that includes income, hours of work, and tenure length. The quality index is then obtained by calculating the first principal component of occupational fixed effects across the seven work characteristics.

Boar and Lashkari (2022) combines data from the 2002, 2006, 2010 and 2014 cycles of the GSS. The dataset includes detailed occupational codes, which Boar and Lashkari (2022) aggregate into 80 categories. To match the GSS occupational codes with ours, we create a correspondance between each detailed occupational code in the GSS and broad (3-digit) occupational categories in the Canadian classification, of which there are 141 (140 when we exclude the "not working" category). Since we estimate the index at a more granular level than in Boar and Lashkari (2022), we include the 2018 and 2022 cycles of the GSS to increase sample size and improve precision.

B Income Mobility and Fields of Study

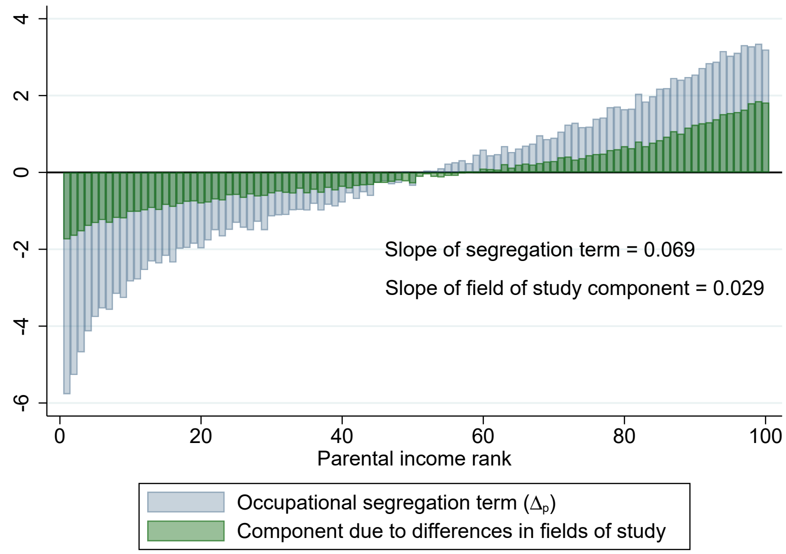
Can occupational sorting be traced back to earlier educational choices? While some occupations require formal degrees in specific fields of study, many do not. That is, the mapping between fields of study and occupations is not deterministic and may vary across parental income groups.

Figure A11, panel A shows that about half of the occupational segregation gradient – the linear slope of Δ_p as a function of parental income rank p – is accounted for by differences in the fields of study children pursue across parental income groups. This result is obtained by assigning each child the value of $\hat{\delta}_o$ for their occupation, and regressing these occupational returns on parental income rank dummies and field of study fixed effects in the micro data. This implies that even conditional on fields of study, children of rich households are systematically more likely to work in high-paying occupations.

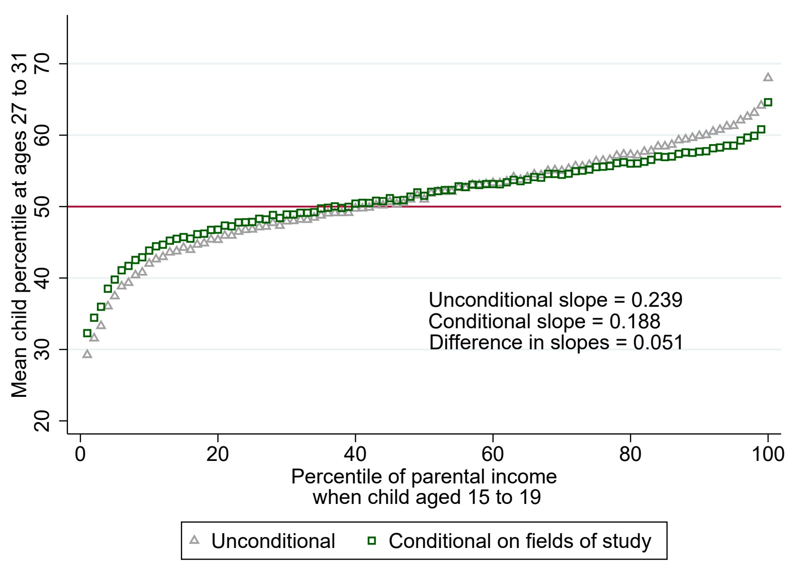
In panel B, we show that fields of study have overall fairly limited explanatory power for intergenerational income mobility. The income rank-rank slope conditional on field of study fixed effects is 0.188, a difference of 0.051 relative to the unconditional rank-rank slope of 0.239. Since fields of study are only observed for individuals with some post-secondary schooling, they also reflect differences in educational attainment. This likely explains why the explanatory power of fields of study appears greater at the top of the parental income distribution.

Figure A11: Intergenerational Income Mobility and Children's Fields of Study

Panel A: Occupational segregation and Fields of Study



Panel B: Child-Parent Income Rank-rank Relationship



Notes: Panel A shows the fraction of occupational segregation by parental income that can be traced back to differences in fields of study. Blue bars show the total contribution of occupations to income mobility, and green bars show the component that is explained by fields of study. To calculate that component, we regress $\hat{\delta}_o$ on parent income rank dummies as well as a full set of field of study fixed effects. The estimated parental income rank coefficients represent the fraction of occupational segregation that operates within fields of study. The green bars are the difference between the total occupational segregation Δ_p and the fraction that operates within fields of study. Panel B replicates Figure 2, substituting field of study fixed effects for occupation fixed effects in equation (1).

C Mechanisms of occupational segregation: An alternative decomposition approach

In this Appendix, we implement an alternative decomposition of patterns of occupational sorting using a Shapley-Owen decomposition in the spirit of Althoff et al. (2023), who estimate the separate contribution of fathers' and mothers' human capital to intergenerational mobility. This method effectively decomposes the R^2 from a multivariate regression into components attributable to each independent variable. It has numerous desirable statistical properties (e.g., additivity of individual contributions) and provides a natural and intuitive summary measure of the contribution of each predictor.

To quantify the contribution of different mechanisms to occupational sorting, we consider the following linear regression of occupational returns $\delta_{o(i)}$ for children i working in occupation o – the dependent variable of interest – on a set of parental characteristics:

$$\hat{\delta}_{o(i)} = \alpha + \mathbf{\Gamma} \boldsymbol{X}_i + \epsilon_i$$

where $\boldsymbol{X}_i$ is a vector of parental characteristics ($\mathcal{X}$ denotes the set of characteristics) and $\hat{\delta}_{o(i)}$ is a measure of occupational returns estimated from equation (1). For consistency, each variable in $\boldsymbol{X}_i$ enters the model as a collection of dummies, as in Section 3. The Shapley-Owen decomposition defines the contribution of each parental characteristic $k \in \mathcal{X}$ to the overall R^2 as

$$\phi_k = \sum_{T \subseteq \mathcal{X} - \{x_k\}} \frac{(n-m-1)!m!}{n!} \left[R^2(T \cup \{x_k\}) - R^2(T) \right]$$

where $R^2(T)$ is the R^2 associated with a regression of $\hat{\delta}_{o(i)}$ on the set of predictor variables $T \subseteq \mathcal{X} - \{x_k\}$, m is the number of predictors in T, and n is total number of predictors in $\mathcal{X}$. The sum is over all possible permutations T of subsets of regressors.[28] One key feature of this approach is that $\sum_{k \in \mathcal{X}} \phi_k$ is equal to $R^2(\mathcal{X})$, that is the R^2 for the full model that includes all predictors.

Note that the approach quantifies the contribution of each factor k to the *total* variation

[28] For example, with 3 regressors x_1, x_2 and x_3, there are 8 permutations, which are $\{\emptyset\}$, $\{x_1\}$, $\{x_2\}$, $\{x_3\}$, $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, and $\{x_1, x_2, x_3\}$. In our setting, we have 8 sets of regressors, for a total of 256 permutations.

in occupational returns. To examine the role of parental characteristics as mechanisms of occupational sorting *by parental income*, we include parental income in $\mathcal{X}$ and implement a series of decompositions for different subsets of predictors.

For reference, the R^2 from a regression that only includes parental income rank dummies in $\boldsymbol{X}_i$ – and therefore pins down the values of Δ_p – is 0.0217. This is the variation we seek to explain. The total R^2 from a regression that includes parental income as well as parental occupations (father's and mother's, separately), parental education (father's and mother's, separately), parental place of birth (father's and mother's, separately), and childhood neighborhood as predictors is 0.0397. In panel A of Figure A12, we report the values of ϕ_k for all 8 predictors, as well as the corresponding R^2 from regressions that only include one predictor. Parental income is the most important contributor to the total variation in occupational returns, followed by childhood neighborhoods. The contribution ϕ_k of parental income is 0.0135, suggesting that other parental characteristics account for $(1 - 0.0135/0.0217 =)$ 38% of the share of the variance in occupational returns that is explained by parental income.[29]

To investigate which parental characteristics contribute the most to patterns of occupational sorting by parental income, let $\phi_{j,-k}$ denote the contribution of factor j for a regression that includes all characteristics except for characteristic k. That is, contribution $\phi_{j,-k}$ of factor j partly includes explanatory power that is attributable to factor k since that factor is not separately accounted for in this model. The total R^2 for the model that includes all characteristics *except for parental income* is 0.0307. That is, adding parental income as a regressor increases the total R^2 by $(0.0397 - 0.0307 =)$ 0.0090 points, hence $(0.0090/0.0135 =)$ 66% of the contribution of parental income is due to independent variation, and a third (or 0.0046 points) is due to shared variation with other predictors.

In panel B of Figure A12, we present the difference between $\phi_{k,-1}$ – the contribution of factor k from a model that excludes parental income (which we index by 1) – and ϕ_k. Intuitively, these differences relate to the importance of the joint variation between characteristic k and parental income. One useful property of the Shapley-Owen decomposition is that the sum of $(\phi_{k,-1} - \phi_k)$ across the 7 candidate characteristics is exactly equal to the component of parental income's contribution that captures joint variation with other factors $(= 0.0046)$. This is because $(\phi_{j,-k} - \phi_j)$ is equal to $(\phi_{k,-j} - \phi_k)$. Hence $(\phi_{k,-1} - \phi_k)$ is the same as $(\phi_{1,-k} - \phi_1)$, which captures how the contribution of parental income changes when

[29] This is somewhat below the combined predictive power of 55% we obtain under our preferred decomposition approach. This is likely because whereas the covariance between parental income and factor k is distributed between each of their contributions in the R^2 approach, our preferred method implicitly allocates all of the joint variation between parental income and factor k to that factor.

factor k is excluded from the model.

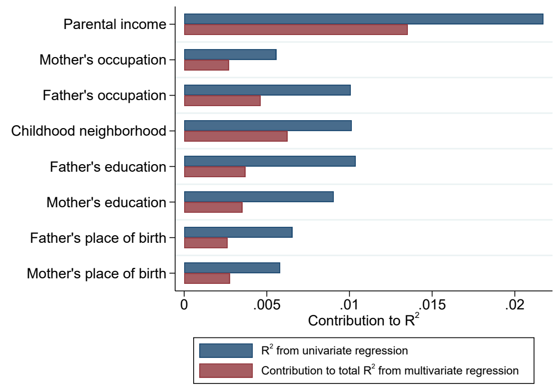
We see that the contribution of mothers' place of birth is largely unchanged when parental income is accounted for, suggesting that its role for occupational sorting by parental income is very small. In contrast, the contribution of paternal occupations is significantly reduced once parental income is accounted for. Interestingly, while the role of childhood neighborhoods for overall patterns of occupational sorting is very important, little of it is related to parental income.

In panel C of Figure A12, we report values of $(\phi_{k,-1} - \phi_k)$. Whereas our preferred decomposition suggested a greater role for parental education (father's and mother's combined) relative to parental occupations, here their importance is similar in good part because we find a somewhat greater role for father's occupation. As in our main results, we find that mothers' education is more important than mothers' occupation, and that parental place of birth does not play an important role. Childhood neighborhood remains less important than both parental education and occupation, but more so than place of birth.

The Shapley-Owen decomposition differs from our preferred decomposition method based on equations (5) and (6) in several ways. First, the Shapley-Owen approach focuses on decomposing the R^2 from a multivariate regression with one continuous outcome. In our setting, this means calculating the contribution of different predictors to the variation in occupational returns, irrespective of what the underlying occupations are (i.e., two different occupations with similar economic returns are treated the same way). Put differently, the method quantifies the extent to which different parental characteristics predict being in a high-paying versus low-paying occupation. In contrast, our preferred method produces counterfactual distributions of child occupations separately for each occupation. By estimating equations (5) and (6) separately for each occupation, our approach allows for the explanatory power of parental characteristics to vary flexibly across each of the 500 child occupations. Second, to examine how patterns of occupational sorting by parental income are explained by other predictors, our adaptation of the R^2 approach relies on residual variation in parental income conditional other parental characteristics. In comparison, our preferred approach implicitly captures joint variation in parental income and other characteristics by taking averages of fitted values (which are based on parental characteristics alone) separately by each parental income rank p. That is, we predict occupations *without* using information on parental income, and then examine the extent to which these predicted distributions of occupations can reproduce the observed patterns of segregation by parental income.

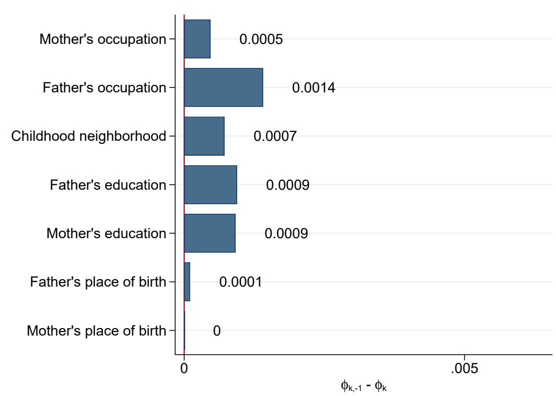
Figure A12: Shapley-Owen Decomposition

Panel A: Full decomposition of R^2



Panel B: Truncated decomposition of R^2 (parental income excluded)

Panel C: Differences in contributions



Notes: This figure presents the results of a Shapley-Owen decomposition of the explained variance in occupational returns. Panel A shows the contributions for a model that includes all 7 characteristics as well as parental income. Panel B shows the contributions for a model that excludes parental income. Panel C shows differences in contributions between the full and truncated models.

D Further patterns of occupational segregation by parental income

The analyses in section 3 quantify the role of different mechanisms for why occupations explain about a third of the income rank-rank relationship. But as indicated before, occupations matter beyond their contribution to income mobility. Here, we report some statistics that summarize the role of different channels of occupational segregation by parental income, without weighting by occupational returns. That is, we calculate counterfactual distributions of children occupations $\tilde{s}_{o|p} = E[\boldsymbol{\beta}^o \boldsymbol{X}_i|p]$, and contrast some statistical properties of these distributions with that of the true distribution $s_{o|p}$.

Since there are over 5,000 values of $\tilde{s}_{o|p}$ (500 occupations across 100 parental income groups), for ease of presentation we calculate separately for each occupation o the variance in occupational shares across parental income groups $Var_o(s_{o|p})$, as well as the variance in the component attributable to parental characteristics $Var_o(\tilde{s}_{o|p})$. We calculate the terms $Var_o(\tilde{s}_{o|p})$ separately for four sets of predictors: parental occupations, parental education, parental place of birth, and childhood neighborhood.[30]

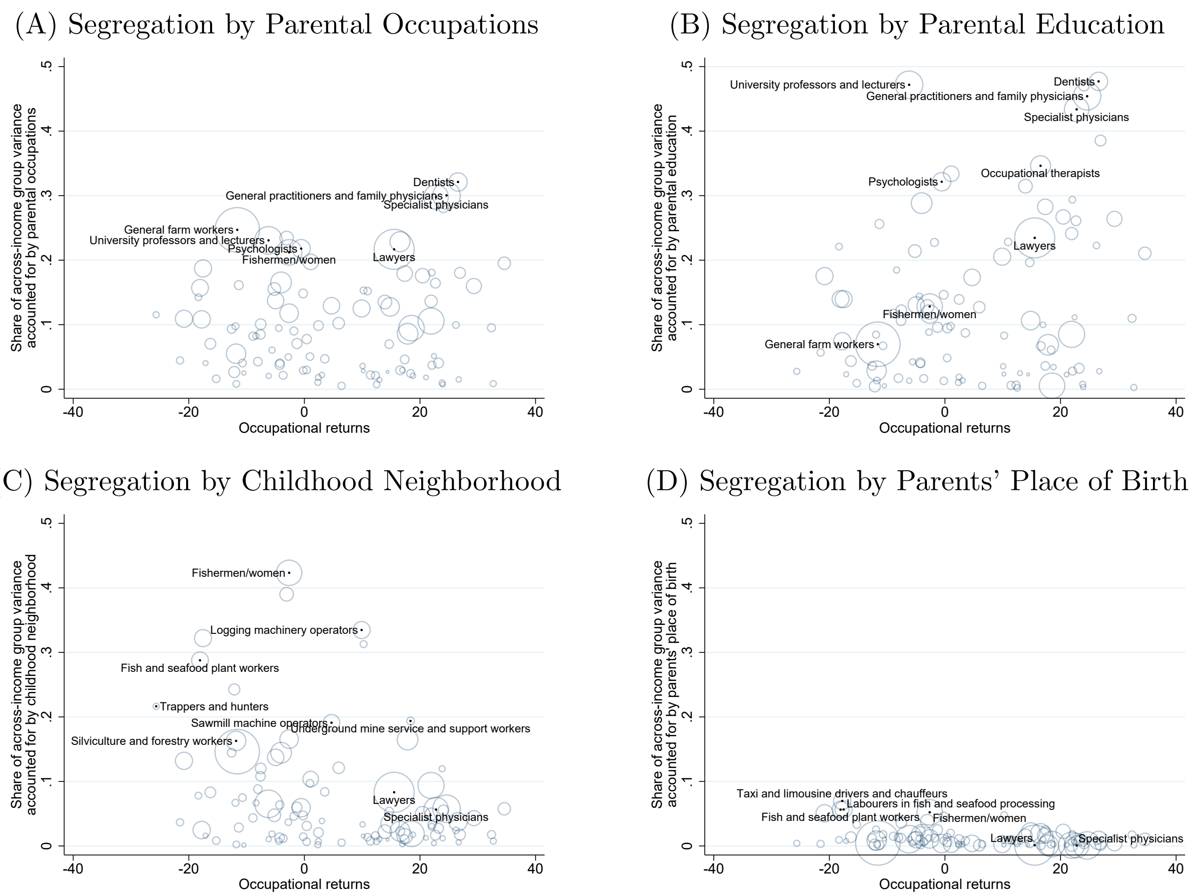
Since the goal is to better understand differences in occupational choice by parental income, we focus on occupations for which the distribution of parental income is considerably uneven (i.e. for which there is significant segregation by parental income to be explained). To do so, we select the top 100 occupations with the largest values of the normalized standard deviation $\frac{\sqrt{Var_o(s_{o|p})}}{s_o}$, a variable that very closely aligns with the Herfindahl index we report on Figure 2. The share of the total variance across parental income groups that is accounted for by a given factor, $\frac{Var_o(\tilde{s}_{o|p})}{Var_o(s_{o|p})}$, is reported on Figure A13 for the top 100 occupations most segregated by parental income.

We highlight selected occupations for which a sizable share of the degree of segregation by parental income can be traced back to differences in parental characteristics. For instance, in panel A, we see that the fact physicians, dentists, and lawyers are highly-segregated occupations is partly due to the fact that selection into these occupations is strongly associated with parental occupations. Strikingly, differences in parental education contribute even more to segregation into these high-paying occupations. In contrast, childhood neighborhoods matter mostly for selection into occupations associated with the agriculture and fishing industries. Finally, differences in parents' place of birth contribute very little to observed patterns of

[30]The $\tilde{s}_{o|p}$ components used here are the same that were used to produce Figure 6, panel C, as well as all three panels of Figure 7.

occupational segregation by parental income.

Figure A13: Mechanisms of Occupational Segregation by Parental Income



Notes: This figure reports the share of the total variance across parental income groups that is accounted for by a given factor, $\frac{Var_o(\tilde{s}_{o|p})}{Var_o(s_{o|p})}$, for the top 100 occupations most segregated by parental income. Panel A considers the role of parental occupations, where panel B focuses on parental education, panel C on childhood neighborhood, and panel D on parents' place of birth. Each circle represents one occupation, and the size of circles is proportional to the number of children in each occupation.