Online Appendix of:

It Runs in the Family:
Occupational Choice and the Allocation of Talent

Mattias Almgren, John Kramer, Jósef Sigurdsson

March 7, 2025

A Additional Analysis and Background Material

A.1 Mapping Swedish Occupational Codes Over Time

The occupational codes in our dataset change over time. Before 1985, occupations are coded according to a three digit code named YK80; between 1985 and 1990, occupations are coded according to YK85, a five digit coding; and after that, occupations are coded according to SSYK96, a three digit coding. In order to facilitate our analysis, we elect to convert all codings into the most current one, SSYK96, at the three digit level.

We obtain a crosswalk between YK85 and SSYK96 from the Swedish statistical office (SCB). Conveniently, the former maps into the latter "m:1", i.e., multiple YK85 occupations map into the same SSYK96 occuption, but not vice versa.

The oldest occupational coding, YK80 also maps into SSYK96, but that mapping is "1:m", implying each of the older occupations maps into multiple recent ones. We tackle this problem by assigning each of the YK80 occupations exactly one SSYK96 counterpart, based on the highest overlap between the two. The tables describing crosswalks between the different occupational codings, produced by the Swedish statistical office, also indicate how many individuals assigned to occupation o in YK80 are assigned to each occupation P in SSYK96. In order to isolate a single SSYK96 occupation to which to assign each YK80 occupation, we pick the one to which most individuals are assigned, separately for men and women. We believe that this creates a credible crosswalk between the two codings. In almost 80 percent of all cases (for men), more than 70 percent of all individuals in a YK80 occupation are coded to one specific SSYK96 occupation and in 60 percent of all cases (for men), more than 90 percent of all individuals in a YK80 occupation are coded to one specific SSYK96 occupation.

A.2 Sensitivity to the Age at Skill Measurement

The data on skills used in this paper are based on measures at age 18. While these measures are intended to capture general skills, they may not reflect innate abilities. Instead, the skills and their measures may be influenced by the environment in various ways. Depending on how quantitatively important such endogeneity is, it could have important implications for our results. Importantly, if fathers invest in the skills of their sons that are most productive in their own occupation, and, in particular, if higher-income fathers engage more in such training than lower-income fathers, we may underestimate the true effect of parental occupation on intergenerational mobility. If skills are endogeneous in this way, we would expect that the relationship between the son's own skills and his father's skills and income would grow stronger over time.

To evaluate this concern, we leverage another source of data where individuals' skills are measured at younger ages. We use data on scores from tests administered as part of the *Evaluation Through Follow-up*, a large survey of Swedish families. These tests are taken when individuals are in 6th grade, at ages 12-13. The data cover around 10 percent of the birth cohorts 1948, 1953, 1967, 1972, and 1977.[40] Härnqvist (2000) and Svensson (2011) provide details on the tests. Importantly, both data sources include tests for logical reasoning and vocabulary knowledge, which were unchanged across the cohorts.[41]

We restrict our sample to individuals for whom we have skills measured in both data sets. Restricting further to individuals for whom we also measure skills of their fathers reduces the sample substantially. We therefore report results both in terms of skills of fathers as well as father's income. We observe the number of questions that each person answered correctly on each test, both in the military enlistment and in the Evaluation Through Follow-up survey, out of a total of 40. We rank individuals by the test score distribution in their cohort. For fathers, we instead aggregate these to decile ranks of skills, due to fewer observations, while using percentile ranks of their income.

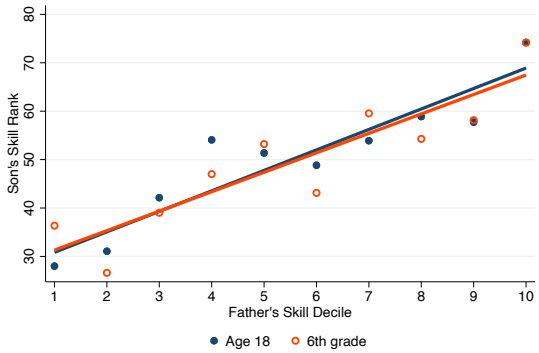
Figure A.1 presents the intergenerational relationships between father's and son's skills, and between father's income and son's skills. Panel (a) plots the relationship between son's and father's logic-inductive ability, at ages 18 and 12/13. Not surprisingly and in line with extensive earlier literature, there is a strong intergenerational correlation of skills. However, this pattern is remarkably similar at both younger and older ages, indicating limited

[40] The sample size of the survey, pooling across all cohorts, is roughly 20,000 individuals.

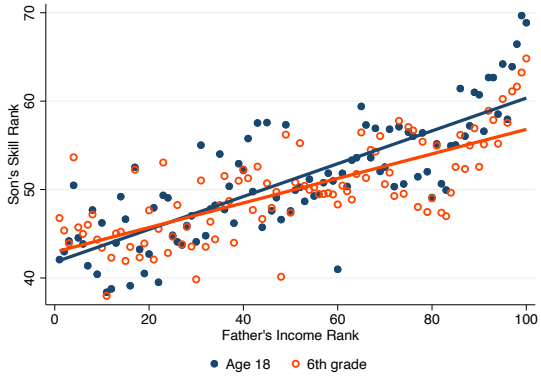
[41] In the *Evaluation Through Follow-up* survey, the test on logical reasoning is to guess a number in a sequence of numbers, and the vocabulary knowledge test is to recognize antonyms (Svensson, 2011). In the military enlistment data, the logical reasoning test consisted of drawing correct conclusions based on statements that are made complex by distracting negations or conditional clauses and numerical operations, and the vocabulary knowledge test consisted of correctly identifying synonyms to a set of words (Carlsted and Mårdberg, 1993).

Figure A.1: Comparison of Skills Measured at Age 18 and Age 12/13

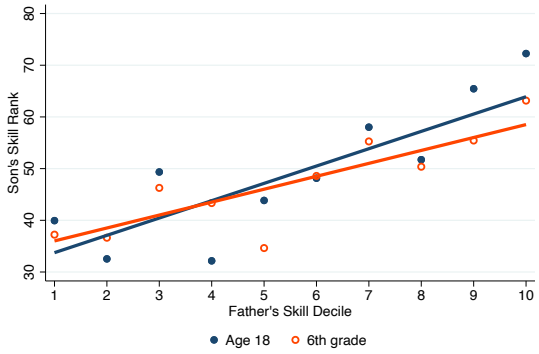
(a) Logic-Inductive Ability by Fathers Skills



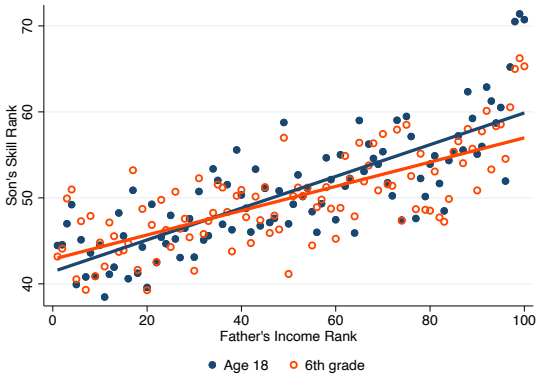
(b) Logic-Inductive Ability by Fathers Income



(c) Verbal Comprehension by Fathers Skill



(d) Verbal Comprehension by Fathers Income



Note: This figure presents the intergenerational relationships between sons' and fathers' skills, and sons' skills and fathers' income rank. Skills are two cognitive skill measures: logic-inductive ability and verbal comprehension. Skills are measured in 6th grade (ages 12/13). The former is based on the *Evaluation Through Follow-up* while the latter is measured in tests administered as part of the military draft. The latter is our main measure used in this paper. Son's skills are measured as the percentile rank in their cohort. Father's skills are measured as a decile in the distribution of fathers within son's cohort, and father's income is measured as percentile rank in the distribution of fathers within son's cohort.

effect of parental skills on their children's skills, above and beyond their initial inheritance. As explained above, the sample size is small where we have the triplet of skills measured at two ages for the sons and skills measured for their fathers. We therefore also present results where we relate skills of sons to income rank of fathers, which we can measure for almost all sons in the sample. As expected, there is a positive relationship between sons' skills and fathers income rank. As with father skills, this relation is almost the same when measured at ages 18 and 12/13. Panels (c) and (d) repeat the same exercise for the case of verbal comprehension, showing similar results.

We conclude from this exercise that we find limited evidence suggesting that skills of sons of high-skilled and high-income fathers change differently than that of lower-skilled

and lower-income fathers over their early lives.

A.3 Family Environment and Brother Comparison

A general concern regarding our methodology is that the relationship between skills and occupational choice may reflect upbringing and the family environment. For example, as highlighted by Becker and Tomes (1986), parents may invest in their child's human capital, e.g., by training them to succeed in their own occupation. Moreover, occupational choice may reflect unobserved skills, perhaps passed on from parents to children.

To evaluate this concern, we study sons in our data have a brother for whom we also have a measure of skills and occupation. If skills are endogenous to parental background, or occupational choice reflects unobserved skills that are common among brothers, this can be differenced out. That is, we can study how the differences in observed skills among brother relates to differences in their propensity to enter occupations.

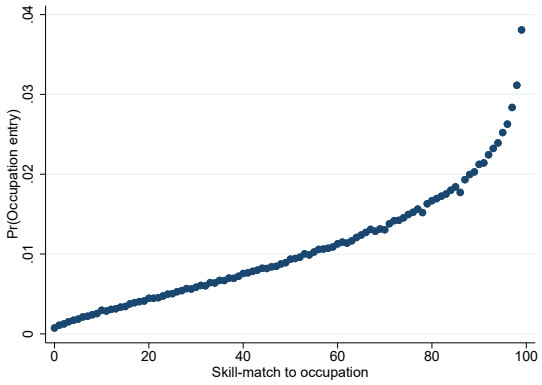
Figure A.2 nonparametrically investigates the relationship between skill-fit, i.e., the entrance probability predicted by our machine learning algorithm, and a son's propensity to choose a given occupation. In all four panels of Figure A.2, skill-fit is plotted on the x-axis. In order for the measure to be comparable across many occupations we generate percentile ranks of probabilities within occupations, such that those with the lowest entry probability have a rank of 1 but those with the highest have a rank of 100. Panel (a) validates our approach by documenting that sons are more likely to enter a given occupation the better their skills match to that occupation. Panel (c) similarly shows that the same is true about occupational following: sons are more likely to enter their fathers' occupation the better their skills match to that occupation. As emphasized above these patterns may both reflect endogeneity of skills to the family environment or skills that are unobserved but important for occupational choice.

To investigate this, panels (b) and (d) first restrict the sample to brothers (blue dots) and then partials out a father fixed effect (orange diamonds). This leaves the relationship between the differences in brothers' skills and their differences in the propensity for occupation entry. If the driver behind the patterns in panels (a) and (c) is family environment, training, or unobserved skills common among brothers, we would expect the line of orange diamonds to be flatter than blue dots. That is, brothers should exhibit a similar propensities for occupation entry. This is not the case. The introduction of fixed effects leaves the slope almost unchanged.

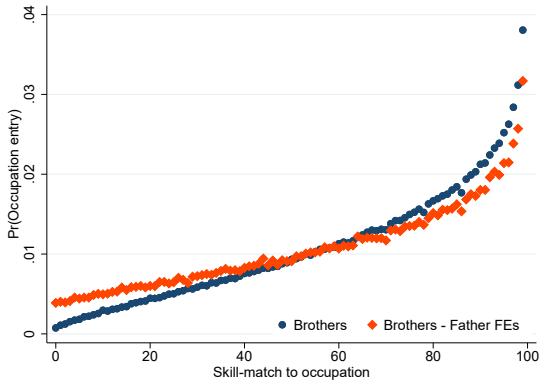
We extend this analysis in Figure A.3 by studying brothers separately by birth order (panel a) and biological and adopted sons separately (panel b). While there is a strong relationship between skill match and following for all sons, first born sons are most likely

Figure A.2: Occupational Choice – Skill Match and Family Background

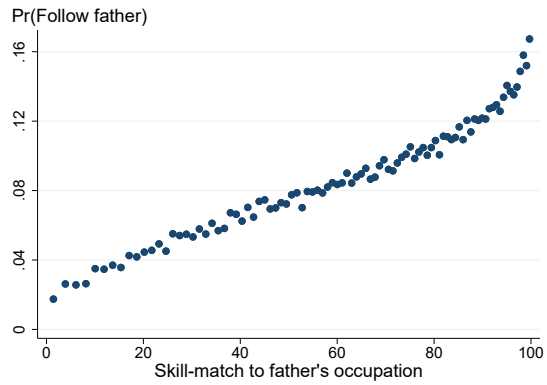
(a) Occupational Choice – All Sons



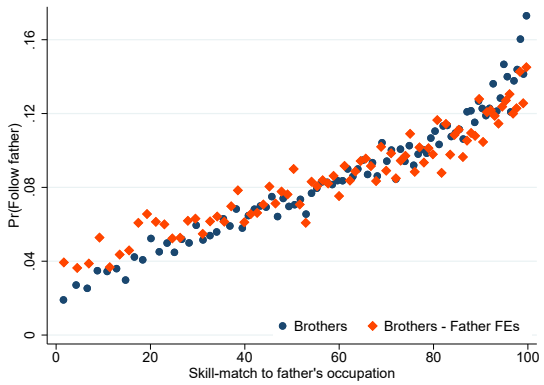
(b) Occupational Choice – Brothers



(c) Occupational Following – All Sons



(d) Occupational Following – Brothers

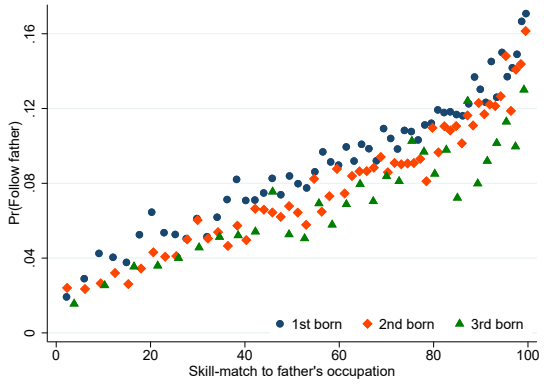


Note: This figure plots binned scatter plots of relationship between (i) the propensity to choose an occupation and (ii) the skill-match to that occupation, measured as the probability of entry predicted based on skills and presented in percentile ranks. All figures are based on regressions that partial out fixed effects for father's occupation. Panel (a) plots the relationship between skill-match and propensity to occupation entry, reflecting the average probability across occupations. Panel (b) plots the relationship between the occupation entry probability and skill match for the sample of sons that have a brother in our sample, where blue dots show the raw relationship and orange diamonds show the relationship in differences across brothers, estimated using a regression including father fixed effect. Panels (c) and (d) plot these relationships restricting to occupations of fathers, i.e. showing the relationship between skill-match and propensity to follow into father's occupation.

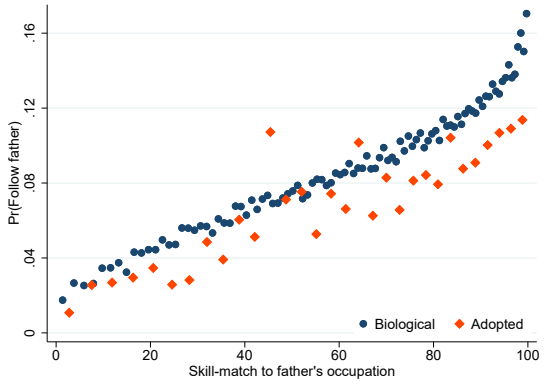
to follow irrespective of skills, roughly 1 percentage points more likely than the second born and 2 percentage points more than the third born. This result speaks to prior studies documenting that earlier born children tend to attain more education (Black et al., 2005), have greater leadership skills, and are more willing to assume responsibility (Black et al., 2018), consistent with parents investing more in earlier than later born children. Lastly, in panel (d) we document that biological sons are 1.4 percent more likely to follow than adopted sons, but we still find a strong skill-gradient of following for both groups.

Figure A.3: Occupational Following – Birth Order and Biology

(a) Brothers by Birth Order



(b) Biological and Adopted Sons



Note: This figure plots binned scatter plots of relationship between (i) the propensity to choose an occupation and (ii) the skill-match to that occupation, measured as the probability of entry predicted based on skills and presented in percentile ranks. The figures are based on regressions that partial out fixed effects for father's occupation. Panel (a) plots the relationship between the propensity to follow and skill match by birth order for the sample of sons that have a brother in our sample. The group of "3rd born" sons includes third and later born sons. Panel (b) plots the relationship between the propensity to follow and skill match for biological and adopted sons.

A.4 Robustness to Approximating Occupation-Specific Skills

A specific concern in our setting is that having a father in a given occupation may mean that his children have skills that are specific to that occupation or result in them developing such skills. If these occupation-specific skills are not captured in the interacted set of the general skills, we might falsely attribute occupational following to heterogeneous entry costs which in fact results from selection on comparative advantage based on unobserved skills.

To address this concern, we proxy for workers' unobserved occupation-specific skills by their father's occupation. That is, for the full population of sons, including followers, we predict their earnings in a given occupation by their general skills as well as father-occupation-specific skills approximated by an indicator of whether his father holds a given occupation. We view this as an important test of the robustness of our results. Adding this proxy significantly improves the accuracy of the prediction. The average R^2 increases from 9.3% in the benchmark model with general skills to 15.7% when adding this proxy for occupation-specific skills.

To evaluate the robustness of our results to this alternative measure of occupation-specific earnings, we estimate our model using these newly predicted earnings and perform the same counterfactual experiment as described in Section 7. Table A.1 summarizes the key model aggregates using this alternative earnings predictions and, for comparison,

Table A.1: Robustness Evaluation of Counterfactual Model Results

	Occupational following	Pr($Q1 \rightarrow Q5$)	Rank-Rank slope	Δ Aggregate earnings
A. Cognitive & non-cognitive skills + proxy for occupation-specific skills				
Baseline	8.4%	9.7%	0.386	-
Counterfactual PE	2.9%	12.5%	0.275	1.8%
Counterfactual GE	3.0%	12.5%	0.277	0%
B. Cognitive & non-cognitive skills				
Baseline	8.4%	9.7%	0.387	-
Counterfactual PE	2.9%	12.6%	0.275	2.0%
Counterfactual GE	3.0%	12.5%	0.278	0.1%

Note: The table presents an evaluation of the robustness of important model aggregates to an alternative prediction of occupation specific earnings. Panel A reports the model estimates and the counterfactual based on a prediction of earnings using cognitive and non-cognitive skills as well as an indicator for having a father in a given occupation as proxy for occupation-specific skills. Panel B reports the same for the benchmark model using cognitive and non-cognitive skills to predict earnings, repeating what is reported in Table 1 in the main text. The table shows aggregates in (i) the baseline economy, (ii) the partial equilibrium economy without parental occupational entry discounts but at baseline prices and (ii) the economy without discounts and general equilibrium prices. The first column shows the percentage of sons who choose the same occupation as their fathers. The second column shows the probability of a son with a father in the first quantile of the father's income distribution moving to the top quantile of the son's income distribution. The third column shows the slope of the relation between the income rank of sons and fathers. The fourth column shows the change in aggregate real earnings from the baseline economy.

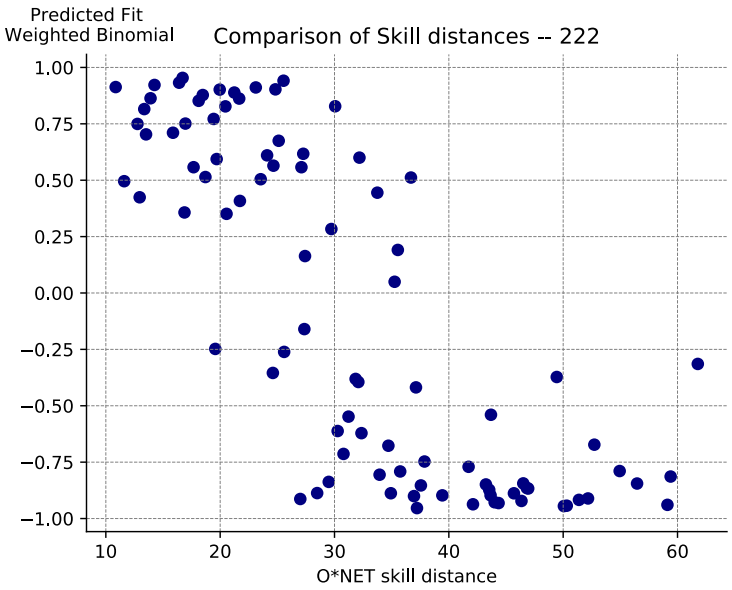
the same aggregates based on our benchmark model. Importantly, we find the results to be robust. The effect of neutralizing entry-cost discounts on occupational following and intergenerational mobility is virtually the same as estimated using our benchmark model. The effect on aggregate earnings is slightly lower, indicating that our benchmark model may attribute some productive skills to unproductive cost discounts. However, these differences are small.

These results suggest that observed occupational following is largely unrelated to occupation-specific skills in father's occupation or other factors that raise earnings of sons in their fathers occupation.

A.5 O^*Net Skill Distance Robustness

As a validation exercise for our ideal occupation predictions, we construct measures of skill distance using them, which can be compared to measures of skill distance calculated using different data.

Figure A.4: Skill Distance and Occupation Similarity for Medical Doctors



Note: This figure shows the skill distance between two occupations, constructed according to Macaluso (2017), using O*NET data, on the x-axis and our measure of occupational similarity on the y-axis. The latter is the outcome of ranking all individuals according to their predicted entry probabilities (i.e., fit probabilities) in two different occupations and then calculating the Spearman correlation coefficient between the two rankings.

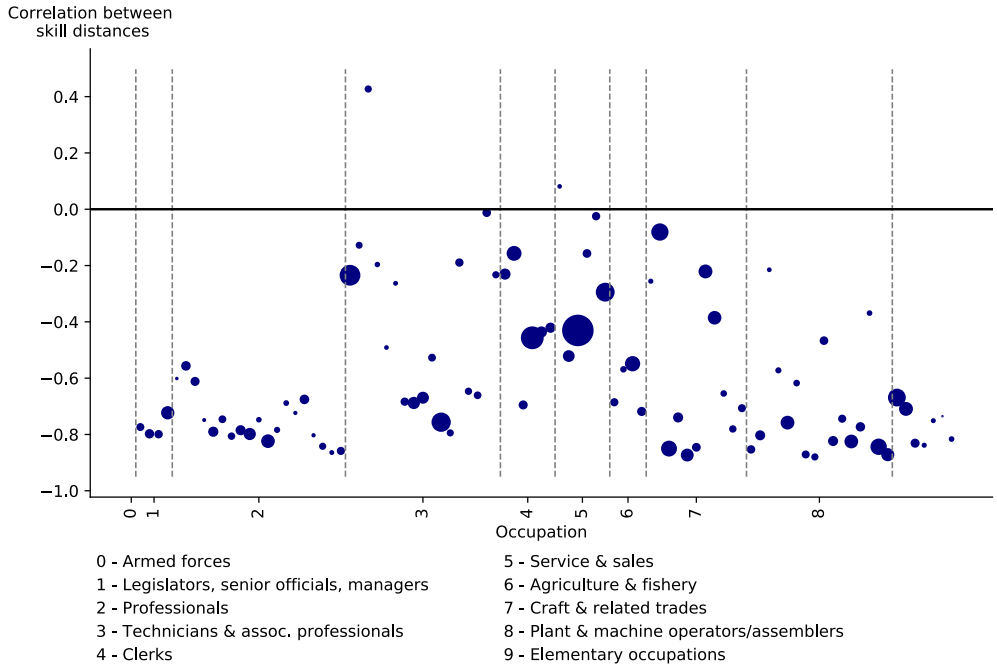
Macaluso (2017) estimates skill distance between two occupations using the O*Net database. Based on surveys, this dataset contains information on the average skillset of incumbents in each occupation, summarized as a 52-dimensional vector. She constructs the distance between occupations as the Manhattan-distance between the two skill vectors in each occupation pair.

First, following her approach, we construct the same measure for our dataset, after mapping the O*NET occupations into Swedish SSYK occuapations, as described in Section A.5.1. Second, to construct the skill distance between two occupations i and j using **our** predictions, we do the following: Using our Random Forest algorithm, we ascertain, occupation-by-occupation, where an individual ranks, in terms of skill fit, *within an occupation*. Using this information, we calculate the Spearman correlation coefficient between the rankings of individuals for every occupation pair i and j in our dataset. If two occupations are more similar, we expect the fit-ranking of individuals to be more similar.

Figure A.4, for medical doctors, shows a clear negative relationship between the skill distance estimated according to the O*NET data (Macaluso, 2017) on the x-axis and our measure of similarity on the y-axis. This gives us some confidence that our random forest algorithm is able to map skill sets into occupations faithfully.

Figure A.5 plots the correlations between the different measures of skill distance across

Figure A.5: Occupational Distance



Note: This figure shows the correlation between two skill distance measures. The first is constructed according to Macaluso (2017), using O*NET data, the second is the outcome of ranking all individuals according to their predicted entry probabilities in two different occupations and then calculating the Spearman correlation coefficient between the two rankings. The y-axis in the figure shows the correlation between the two measures. On the x-axis, ccupations are ordered according to their 3-digit code in the SSYK-96 classification system, the vertical and horizontal lines mark the borders of 1-digit occupational groups.

all occupations.[42] It is negative in almost all cases. The two approaches seem most consistent for the occupations including legislators and professionals, groups 1 and 2. Towards the blue collar occupations, while still negative, the two measures correlate less clearly.

A.5.1 Mapping International Occupational Codes into Swedish Codes

The O*NET database classifies occupations according to an SOC code. In order to map these into the Swedish SSYK96 system, we first map the SOC2010 code into an ISCO-08 code, which can then be mapped into SSYK2012, and finally into SSYK96.

The mapping between SOC2010 and the four-digit ISCO-08 classification is many-to-many. To calculate an ISCO-08 occupation's intensity in each of the different 52 different skills contained in the ONet database, we take the average of each of the skill measures across all SOC2010 occupations that map into it. For hypothetical ISCO occupation $I-1$, we first find all SOC occupations that are linked to it, e.g., hypothetical occupations $S-1$ and $S-2$. To calculate the "oral comprehension" intensity of the $I-1$ occupation, we take

[42] Note that the O*NET database contains no information on military occupations.

an average of the intensity in that skill across $S-1$ and $S-2$, weighted by the employment shares in $S-1$ and $S-2$.[43]. We proceed the same way for all other skills, e.g., "written comprehension" etc; and all other ISCO-08 occupations. Having done this, we obtain a dataframe containing the skill intensity for each of the ISCO-08 occupations, and all skills measured in the ONet database.

ISCO-08, in turn, maps into SSYK12 many-to-many. We use the same approach as before. First, to each SSYK12 occupation, we match all the ISCO-08 occupations that are linked to it. Then, we take the average over all the ISCO-08 occupations within each SSYK12 occupation, by skill. Thus, we obtain a dataframe containing the skill intensity for each of the SSYK12 occupations, and all skills measured in the ONET database.

From SSYK12 we proceed as in step one: merging SSYK12 to SSYK96 occupations and then obtaining average skill intensities for each skill-occupation pair by taking weighted averages, by SSYK12 occupation size.

A.6 Skill Distance

Recall that each individual in our model has mass 1 which is potentially distributed across all occupations (due to the preference shocks). Thus, when moving from the baseline to the counterfactual economy, occupational changes do not occur discretely, i.e., from one occupation to another, but rather as a change in an individual's mass distribution across occupations. To quantify the distance between these distributions, we proceed in two steps. First, we take their difference, to determine how much mass is shifted. We assume that all *reductions* in mass allocated to occupations (when moving from baseline to counterfactual) are distributed randomly to those occupations which *gain* mass in the counterfactual economy. Then, in the second step, we determine how far, on average, the mass lost in each occupation travels to the new occupations. We take the average of these distances, within each sender occupation, across all receiver occupations, weighted by the share mass received in each occupation. This procedure quantifies, for each sender occupation, the average distance to receiver occupations. The final step is to average these distances across sender occupations, weighted by the share of total mass sent. This gives us, at the individual level, the average distance the shifted mass traverses when moving from baseline to counterfactual economy.

[43]We obtain employment shares for all SOC occupations in 2014 from the BLS https://www.bls.gov/oes/tables.htm

A.7 Earnings Measure and The Extensive Margin

In our main analysis, we measure earnings as the wage earnings in worker's primary job and measure occupation as the occupation of that job. For this we use administrative data from the salary structure statistics (Lönestrukturstatistiken), which contains data sampled from firms. Every year, this data includes half of firms in the private sector, sampled at random, and all of the public sector. As income and occupation is measured at prime age, defined as age 30-40, most individuals in our cohort are observed at least several years in this data.

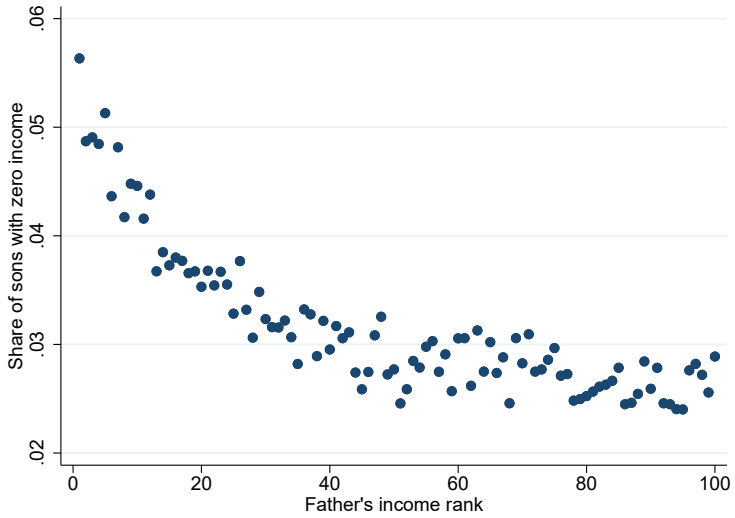
Importantly, with this measure of wage earnings, we restrict our earnings measure to include those that are employed. In addition, we measure earnings in full-time equivalent terms, meaning that apply the monthly salary also to months when workers are not working, e.g. due parental leave, sickness absence, unemployment etc. We argue that this measure is preferable to average annual earnings for two reasons. First, we consistently measure earnings associated with work in a given occupation. Second, this measure of labor income is closer to a measure of wage than earnings, which is preferable when measuring the returns to skills, as we do when measuring potential earnings across occupations. In addition, since our analysis attaches a single prime age occupation to each individual and focuses on occupational choice, it is not trivial to incorporate decisions about the extensive margin into our analysis.

To evaluate this decision, we have also carried out analysis using a measure of total annual labor earnings according to tax records. Although this has some effects on the measured earnings, our main results, such as the difference in the association between son's and father's earnings ranks between the baseline and the counterfactual are broadly similar to when using our preferred earnings measure.

The key difference between the two earnings measures is the extensive margin. We recognize that this has implications for our measures of intergenerational income mobility to the extent that it captures differences in employment. Figure A.6 shows the resulting relationship between the sons' average non-employment incidence and their fathers' income ranks. As before we measure earnings during ages 30-40. For each individual, we compute the fraction of times that we do not observe an income. Since the cohorts that we consider in this exercise are born between years 1950 and 1979, they will be active in the labor market at different points in time, and therefore, simply because of timing, be more of less exposed to periods with high or low aggregate unemployment rates.[44] For this reason, we partial out cohort fixed effects from the aforementioned fractions.

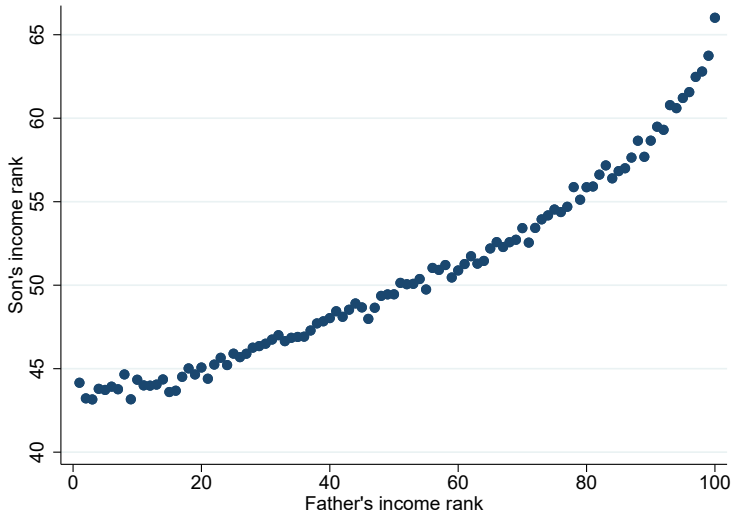
[44]The cohort that was born in year 1960 will be in the prime age bracket between years 1990-2000. During many of those years, the unemployment rate in Sweden was unusually high.

Figure A.6: Share of Individuals with Zero Earnings



Note: This figure shows the share of individuals between the ages 30 and 40 who do not have an income observation. The fraction is adjusted for cohort fixed-effects. The sample period is 1985-2013.

Figure A.7: Association between Son's and Father's Incomes



Note: The figure shows the relationship between son's and their fathers' income ranks. Income is measured as total taxable income according to tax records. Income of sons is measure as the average income at ages 30-32 and income of fathers is measured as the average income at ages 45-47. The sample period is 1985-2013. Zero income is given a rank of zero. Fathers are placed into 100 percentile bins. For each such bin, we calculate the average income rank of sons, which is then plotted on the y-axis. Fathers and sons are ranked within cohort-year cells. The rank-rank slope, estimated wit OLS regression, is 0.190 (SE 0.005)

Figure A.6 documents that the non-employment share is declining with fathers income. This has implications for the shape of the rank-rank association. Prior work has documented that measuring intergenerational mobility in terms of income ranks is useful as the relationship is linear in ranks (Chetty et al., 2014). Our figure 3 shows a pattern that is slightly convex. This reflects the differences in the non-employment incidence by father income rank. Figure A.7 plots the rank-rank association using a measure based on total labor earnings, including the extensive margin, where those with zero earnings are given the lowest rank. As the figure displays, the association is more linear when accounting for non-employment.

B Prediction of Potential Earnings and Occupation Entry

An important input in our structural model and our empirical analysis more generally are measures of potential earnings that an individual would have across all occupations depending on his skills, and similarly the likelihood of entry into occupations (i.e. skill match). To this end, we take a machine-learning approach, where we use a random forest algorithm to flexibly use individuals' skill sets to predict earnings and entry probability.

B.1 Data Preparation for Predicting Earnings

As the prediction is carried out sequentially by occupation we prepare for each occupation two data sets: training data and test data. The former includes all the incumbents in the occupation, excluding those that have fathers with the same occupation (*followers*). We tune the parameters of the algorithm (e.g. depth of trees, learning rate, etc) by drawing a random 10% sample from the training data and predicting for the remaining 90%. Once the algorithm is tuned, train the model on the training data and predict for everyone (test data). This gives us predicted earnings for every individual in all possible occupations.

The prediction is based on residualized income in logs. That is, we estimate the following regression:

$$\ln(earn_i) = \rho_o + \delta_c + \gamma_y + \varepsilon_i$$

where ρ_o, δ_c, and γ_y are, respectively occupation, birth cohort, and calendar year fixed effects. Then we use our machine learning approach to predict the earnings residuals across individuals and occupations. When translating the earnings predictions into SEK, we add fixed effects from the aforementioned regression. For comparability across the sample of individuals, we normalize earnings within each occupation by age and time, such that the

reference age is 40 in a period. We split our sample into six periods, two per decade. Children are assigned to the period in which we observe their prime income, i.e., the income in their modal occupation between ages 30 and 40. The six periods are: 1985, 1990, 1996-1999, 2000-2004, 2005-2009, 2010-2013.

B.2 Data Preparation for Predicting Probabilities

The procedure for predicting entry probabilities is analogous to the procedure for predicting earnings, except for the fact that the prediction is binary as opposed to linear. The test sample is as for the earnings prediction. The training sample adds another restriction as we restrict incumbents to the top 20% earners in each occupation. Our results are not very sensitive to changing the size of this group.

B.3 Prediction

For each individual and each possible occupation, we predict the potential earnings and the probability that the individual takes up that occupation based only on his skills. Training the algorithm on the skills of incumbents in each occupation, this results in a measure of returns to skills in a given occupation (potential earnings) and a measure of how well individuals fit into a given occupation (entry probability). To account for the fact that occupations vary a lot in size, which will influence how accurately we can predict probabilities for small occupation, we use occupational-size weights in the model estimation.

The prediction process is a Random Forest estimation with cross validation (i.e. out-of-sample testing). The Random Forest algorithm is standard, where the number of splits are penalized if they do not yield a sufficient increase in prediction power.[45] The cross-validation procedure works as follows:

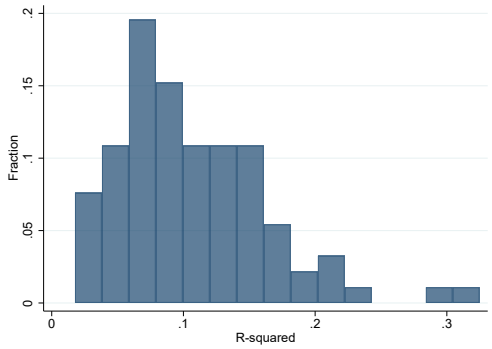
1. The dataset X is split into n subsamples, $X_1, X_2...X_n$.

2. The XGBoost algorithm fits a boosted tree to a training dataset comprising $X_1, X_2, ..., X_{n-1}$, while the last subsample, X_n is held back as a validation (out-of-sample) dataset. The chosen evaluation metrics (RMSE) are calculated for both the training and validation dataset and retained.

3. One subsample in the training dataset is now swapped with the validation subsample, so the training dataset now comprises $X_1, X_2...X_{n-2}, X_n$, and the validation (out-of-sample) dataset is now X_{n-1}. Once again, the algorithm fits a boosted tree to the training data, calculates the evaluation metrics and so on.

[45] We use the XGBoost package in **R**.

Figure A.8: R^2 across Occupation-Level Predictions



Note: This figure plots the distribution of R^2 from random-forest predictions in each occupation. Prediction is based on the eight cognitive and non-cognitive skills of incumbents in each occupation. The sample period is 1985-2013.

4. This process repeats n times until every subsample has served both as a part of the training set and as a validation set.

5. Now, another boosted tree is added and the process outlined in steps 2-4 is repeated. This continues until the total number of boosted trees being fitted to the training data is equal to the number of rounds (i.e. the forest size).

6. There are now n calculated evaluation scores for each round for both the training sets and the validation sets. The prediction is then based on the round that best satisfies the evaluation metric (minimizes RMSE).

Based on the resulting model for a given occupation, we then construct predicted earnings (or entry probabilities) for all individuals. The same procedure is then carried out for all occupations. Figure plots the histogram of R^2 across all occupation-level predictions of earnings. The average R^2 is 0.093.

C Computation Appendix

C.1 Calibration of Baseline Economy

As described in Section 5.3, the baseline economy is calibrated to match data moments related to occupational choices. Costs and discounts are estimated jointly, as each of them affects all model moments. When we estimate the model, we do so in utility terms:

$$u(i,k) = \frac{Y(x(i),k)}{P} - b_k^f \tag{15}$$

where $Y(x(i),k)$ is the nominal income (and nominal expenditure) of individual i who works in occupation k, and P is the aggregate price index in the economy. b_k^f is the utility cost for entering occupation k, when individual i's father is in occupation f. See Equation (12) for more details.

We find initial guesses for our solution method as follows:

1) We consider entry costs only and target the share of sons in different occupations. The entry cost into military occupations is normalized to zero. Once we find an entry cost vector that yields shares that closely align with the corresponding data moments, we stop and store the vector as $m^{0,1}$.

2) Next, we target the shares of sons who choose the same occupational type (blue collar/white collar) as their fathers, taking $m^{0,1}$ as given. We iterate until we find that the model moments are close to their corresponding data moments. Call the resulting vector $d_1^{0,1}$. We normalize the discount for choosing a white-collar occupation to zero. This requires adjustments to the blue collar discounts and the entry cost vector, in order to keep incentives the same. Label the adjusted vectors m^0 and d_1^0, respectively.

3) In the next step, we take take m^0 and d_1^0 as given and search for a vector of one-digit following discounts that brings the model close to the data. Once the model matches the data in this dimension, we store the resulting vector and call it d_2^0.

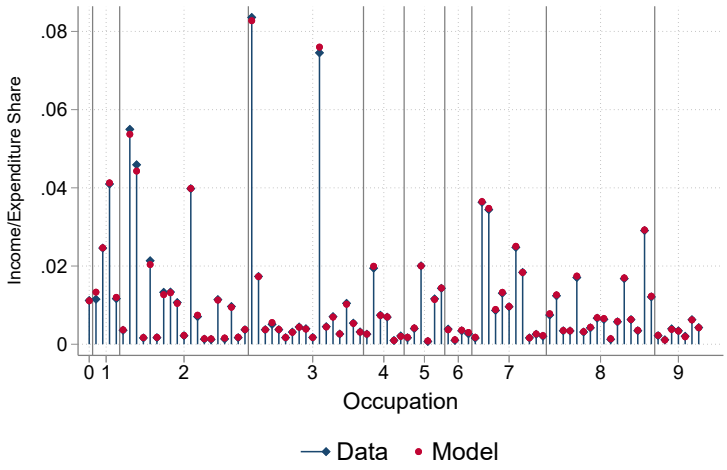
4) Last, we find a first guess for the set of follower discounts, holding all other discounts and costs fixed. We call this vector d_3^0. We normalize the follower discount into armed forces to zero.

Next, we iterate on all costs and discounts simultaneously, starting with the initial guesses obtained according to the above procedure, until the model moments match the data moments that we target. The estimated vectors are $m,\ d_1,\ d_2$, and d_3.

C.2 Counterfactual

In the counterfactual economy, we remove all discounts related to occupational following, and, following the use of the Cobb-Douglas aggregator for preferences, target the expenditure shares at their baseline values. To clear product markets, all prices $\{P_n\}_{n=1}^N$ adjust. For the baseline economy, we assumed that $P_n = 1\ \forall n$. As mentioned in Section 5.3, this normalization has no effect on relative predicted earnings across individuals within occupations, which is what matters for the results in the baseline economy. To find a new price vector $\{P_n^c\}_{n=1}^N$, given the entry costs m, estimated productivities $Z(x,n)$, and expenditure

Figure A.9: Expenditure shares—Data and Baseline model



Note: This figure shows the fraction of income accruing to each three-digit occupational group in the data (blue diamons) and the model (red circles). On the x-axis, occupations are ordered according to their 3-digit code in the SSYK-96 classification system, the horizontal lines mark the borders of 1-digit occupational groups.

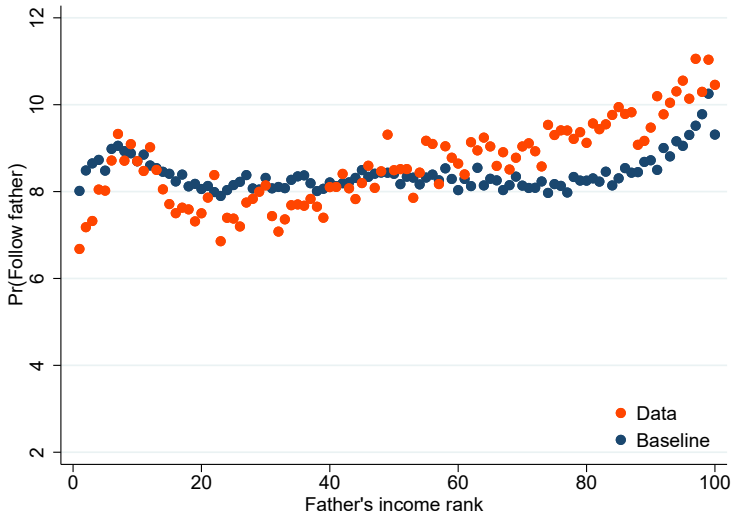
shares $\{\alpha_n\}_{n=1}^N$, we iterate on the price vector until the expenditure shares converge to the data values. As entry costs are measured in utils, we transform income to consumption utility by deflating nominal earnings by the price index $P^c = \prod_n \left(\frac{P_n^c}{\alpha_n}\right)^{\alpha_n}$, like in Equation (15).

D Untargeted Moments

When estimating the occupational-specific entry costs and following discounts, we target the occupational densities, i.e., the share of individuals in each occupation, and parts of the occupational transition matrix between fathers and sons. Encouragingly, the model replicates other, untargeted, features of the data well. First, and most importantly, the model is able to match the expenditure shares across different occupations (which are equivalent to income shares). Figure A.9 shows that although our estimation only targets the share of *individuals*, the model replicates the corresponding shares of *incomes*. This is not a mechanical relationship, and implies that the model reproduces a similar average skill level in each occupation as we see in the data.

Secondly, we document that the model is able to replicate the propensity to follow over the father's earnings distribution. Figure A.10 plots occupational following in the data and in the baseline economy, by father's earnings rank. In general, the model is able to capture both the level of following as well as the differences in following by background.

Figure A.10: Occupational Following in Data vs. Model Baseline

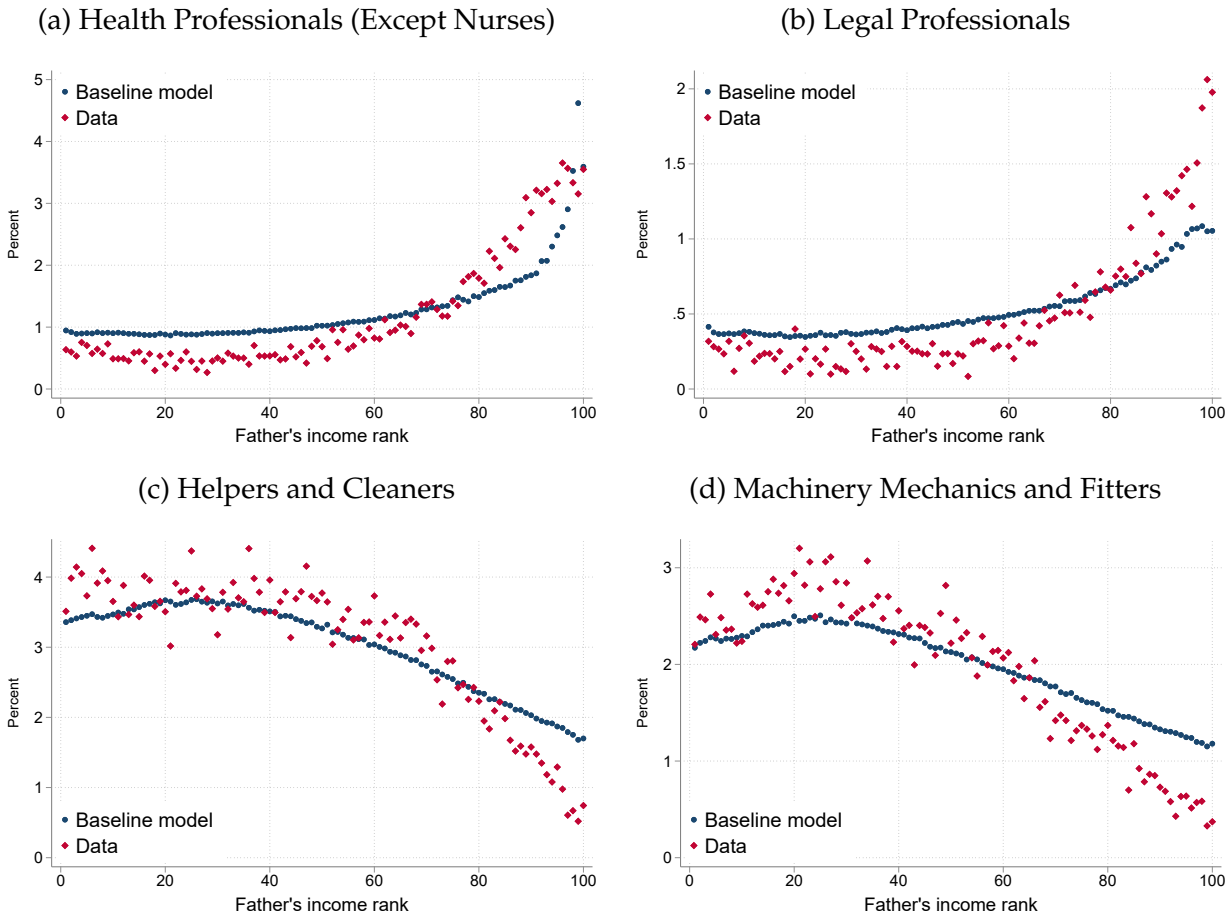


Note: This figure shows share of workers following into the occupation of their father in the data and the propensity for occupational following in the baseline model. The figure plots averages by fathers income percentile rank.

Finally, Figure A.11 shows the shares of children who choose four different occupations, sorted by their fathers' income ranks, comparing the model to the data. Importantly, followers, a fraction which was explicitly targeted in our calibration, are excluded from these graphs. The data shows that individuals born to fathers at the top of the income distribution are close to three times more likely to become health or legal professionals than sons born to fathers at the low end of the income distribution. Conversely, the children of low-earning fathers are much more likely to choose to become cleaners or mechanics than children of high-earning fathers. The model replicates these patterns fairly closely.

E Supplementary Figures and Tables

Figure A.11: Occupational Choice by Father's Income Rank



Note: These figures plot the shares of individuals who choose four different occupations, depending on their fathers' income ranks. All figures exclude sons who choose the same occupation as their father, i.e., occupational followers. The blue dots represent the shares in the data; the red diamonds represent the shares in the calibrated baseline model. Panel (a) plots the share of sons who become health professionals, panel (b) plots the share of sons who choose to become legal professionals, panel (c) plots the share of sons who become helpers and cleaners, and panel (d) plots the share of sons who become mechanics and fitters. The sample period is 1985-2013.

Figure A.12: Mobility Bias across Occupations – Mothers and Daughters

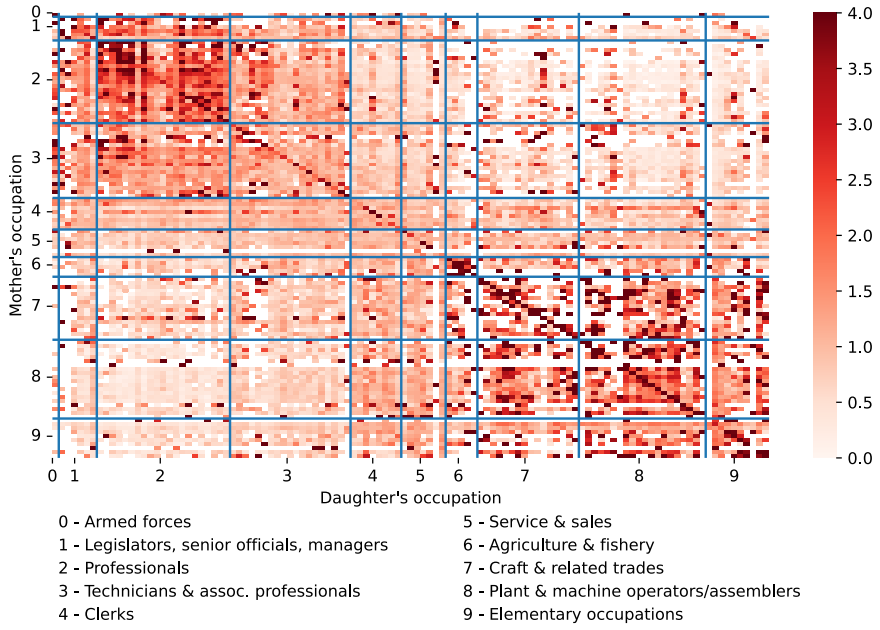


Figure A.13: Mobility Bias across Occupations – Mothers and Sons

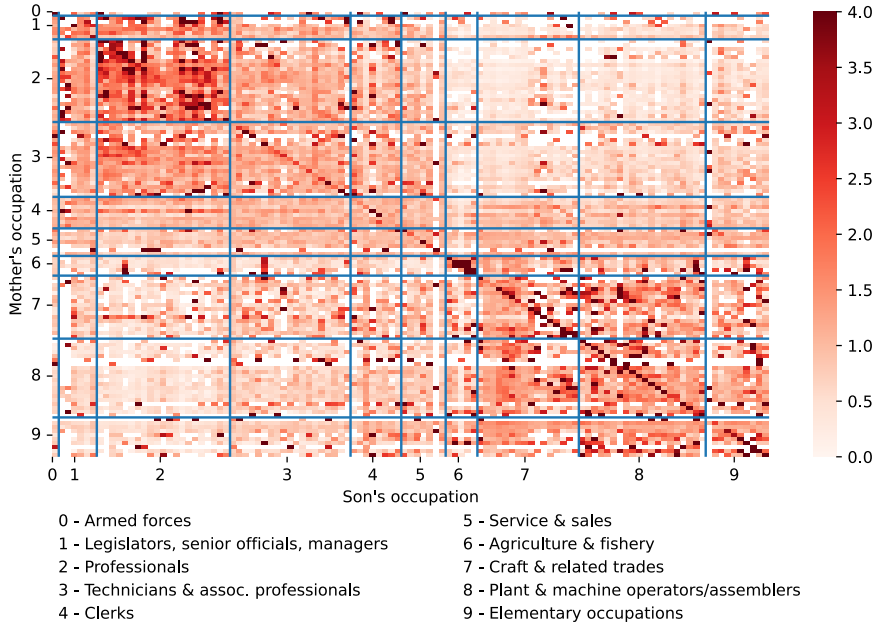


Figure A.14: Mobility Bias across Occupations – Fathers and Daughters

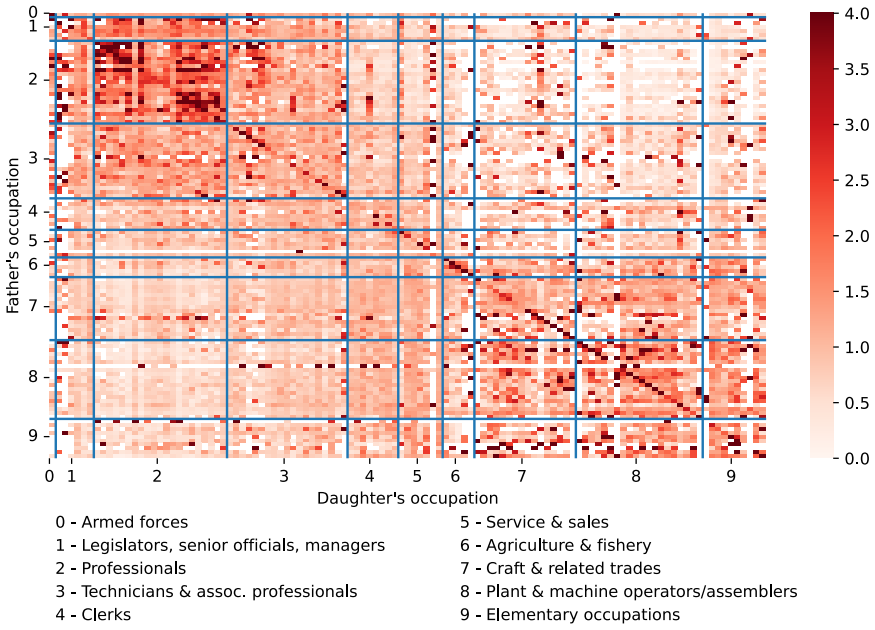
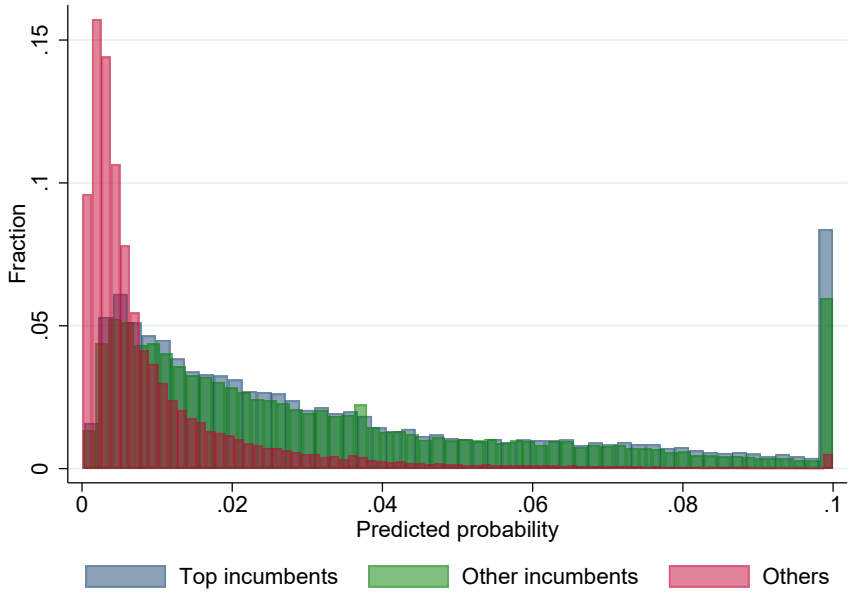
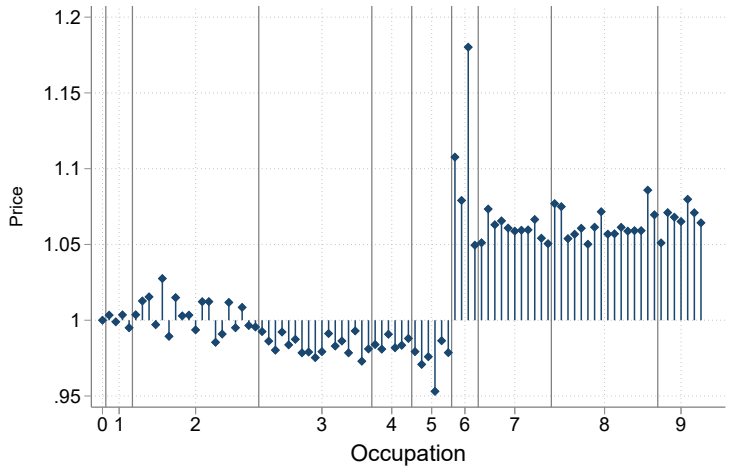


Figure A.15: Predicted Probability of Occupation Entry



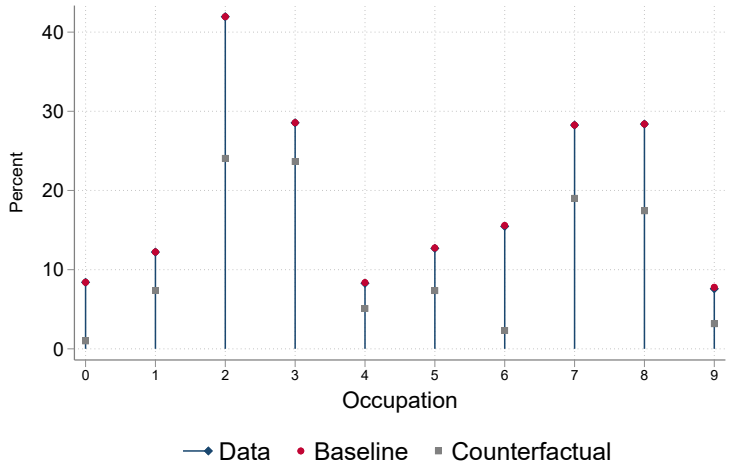
Note: The figure shows predicted probability of entry into occupations. The figure separated three groups: "Top incumbents" which are incumbents in the occupation in the top quintile of the earnings distribution and those used for training the machine-learning algorithm, "Other incumbents" which includes all other incumbents in the occupation, and "Others" which are workers in other occupations. The figure is winterized from above at 10 percent probability of entry.

Figure A.16: Price changes in General Equilibrium



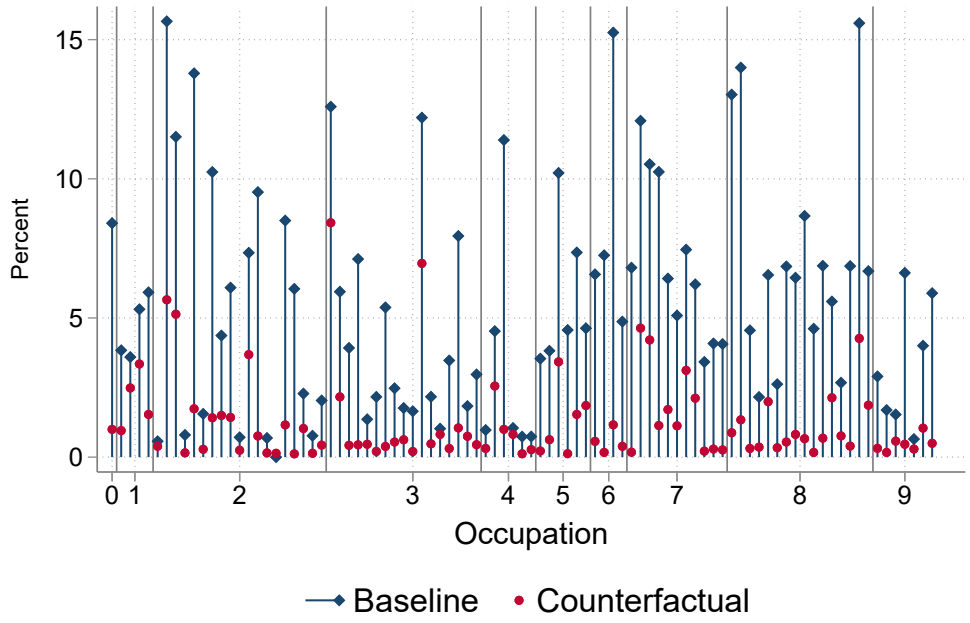
Note: This figure shows the change in the prices for goods produced in each of 91 occupations in the counterfactual economy. Prices in the baseline economy are normalized to one, as is the price for military occupations occupations in general equilibrium (occupational group zero). Occupations are ordered according to their 3-digit code in the SSYK-96 classification system, the vertical and horizontal lines mark the borders of 1-digit occupational groups. For the definition of the mobility bias, see the text. The sample period is 1960-2013.

Figure A.17: Single Digit Occupational Following – Data, model, and counterfactual



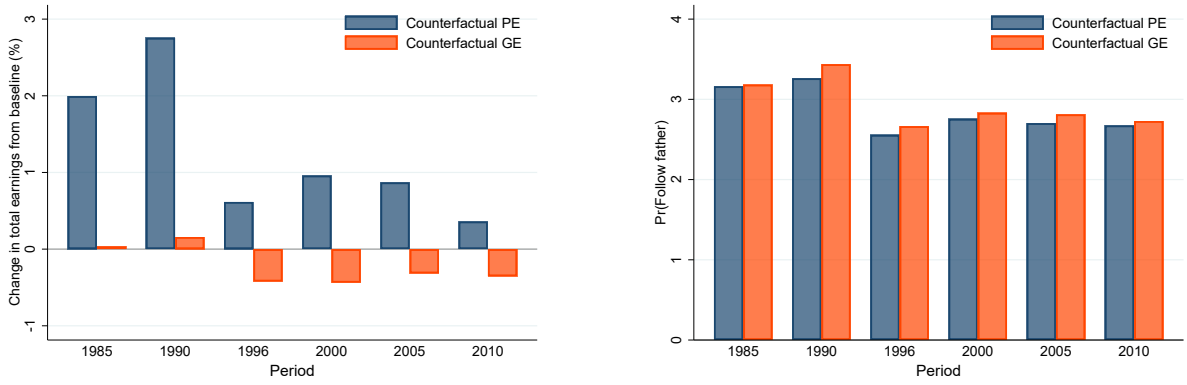
Note: This figure shows the fraction of fathers whose child follows them into the same broad occupational category, i.e., one-digit occupational classification. The blue diamonds represent this fraction for the pooled dataset, the red circles report the results for the baseline model and the gray squares report the results from our counterfactual exercise (see text). On the x-axis, occupations are ordered according to their 3-digit code in the SSYK-96 classification system, the horizontal lines mark the borders of 1-digit occupational groups. The sample period is 1985-2013.

Figure A.18: Following in the Counterfactual Economy



Note: This figure shows the fraction of fathers whose child follows them into the same occupations, for each occupation. The blue diamonds represent this fraction for the baseline model, the red circles report results for the counterfactual economy. On the x-axis, occupations are ordered according to their 3-digit code in the SSYK-96 classification system, the horizontal lines mark the borders of 1-digit occupational groups.

Figure A.19: Aggregate Earnings and Following in Counterfactual Economy by Period

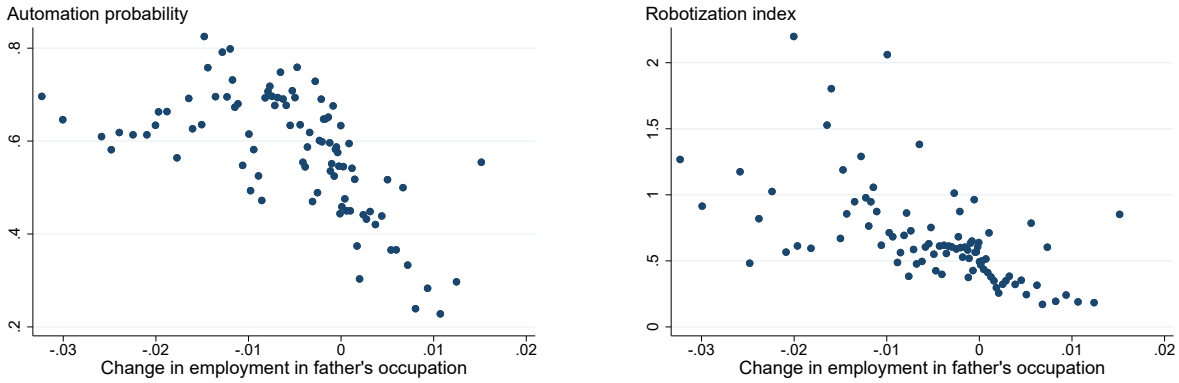


(a) Aggregate Earnings

(b) Occupational Following

Note: The figure plots the partial equilibrium (blue) and general equilibrium (orange) effects of removing all entry cost discounts from the model for each of our six model sub-periods. Panel (a) plots the effects on aggregate earnings, measured as the percentage change in total real earnings relative to the baseline economy. Panel (b) plots occupational following probability. The six periods are: 1985, 1990, 1996-1999, 2000-2004, 2005-2009, 2010-2013.

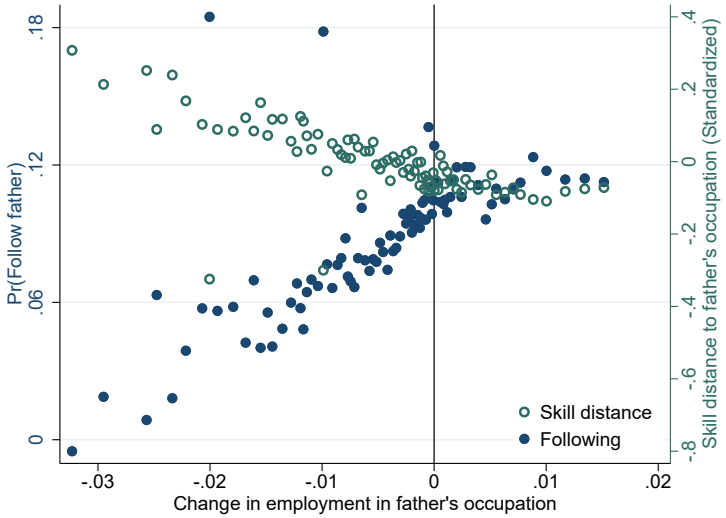
Figure A.20: Occupational Decline: Automation and Robotization



(a) Automation (b) Robotization

Note: The figure plots a binned scatter of the correlation between (i) a change in employment share in fathers' occupation for all sons in our sample and (ii) two measures of labor-saving technological change. In panel (a) we plot occupation-specific automation probabilities based on Frey and Osborne (2017). This measure is based on analysis of 702 US occupation and measures probability in 2010 that an occupation will disappear within 10-20 years due to computerization. Using this measure, Gardberg et al. (2020) also document a decline in employment share since the 1990s in occupations more exposed to risk of automation. In panel (b) we plot occupation-specific measure of exposure to automation measured by tasks that can be performed by industrial robots, as measured by Webb (2019).

Figure A.21: Effect of Employment Decline in Father's Occupation on Skill Distance



Note: The figure plots the relationship between (i) the change in employment share in fathers' occupation since prime age and (ii) both the propensity of sons following into same occupation as their father (left axis) and the occupational skill distance to father's occupation (right axis). The figure is a graphical representation of difference-in-differences regression (13) as it plots a binned scatter plot controlling for occupation and year-at-prime-age (cohort) fixed effects, as well as demographic controls including sibling indicator, and birth order dummies.

Table A.2: List of Occupations: SSYK-96 Codes and their Descriptions

SSYK96 code	Description
011	Armed forces
121	Directors and chief executives
122	Production and operations managers
123	Other specialist managers
131	Managers of small enterprises
211	Physicists, chemists and related professionals
213	Computing professionals
214	Architects, engineers and related professionals
221	Life science professionals
222	Health professionals (except nursing)
223	Nursing and midwifery professionals
231	College, university and higher education teaching professionals
232	Secondary education teaching professionals
233	Primary education teaching professionals
235	Other teaching professionals
241	Business professionals
242	Legal professionals
243	Archivists, librarians and related information professionals
244	Social science and linguistic professionals (except social work professionals)
245	Writers and creative or performing artists
246	Religious professionals
247	Public service administrative professionals
248	Administrative professionals of special-interest organisations
249	Psychologists, social work and related professionals
311	Physical and engineering science technicians
312	Computer associate professionals
313	Optical and electronic equipment operators
314	Ship and aircraft controllers and technicians
315	Safety and quality inspectors
321	Agronomy and forestry technicians
322	Health associate professionals (except nursing)
323	Nursing associate professionals
331	Pre-primary education teaching associate professionals
332	Other teaching associate professionals
341	Finance and sales associate professionals
342	Business services agents and trade brokers
343	Administrative associate professionals
344	Customs, tax and related government associate professionals
345	Police officers and detectives
346	Social work associate professionals
347	Artistic, entertainment and sports associate professionals
412	Numerical clerks
413	Stores and transport clerks
415	Mail carriers and sorting clerks
419	Other office clerks
421	Cashiers, tellers and related clerks
422	Client information clerks
511	Travel attendants and related workers
512	Housekeeping and restaurant services workers
513	Personal care and related workers
514	Other personal services workers
515	Protective services workers
522	Shop and stall salespersons and demonstrators
611	Market gardeners and crop growers
612	Animal producers and related workers
613	Crop and animal producers
614	Forestry and related workers
711	Miners, shotfirers, stone cutters and carvers
712	Building frame and related trades workers
713	Building finishers and related trades workers
714	Painters, building structure cleaners and related trades workers
721	Metal moulders, welders, sheet-metal workers, structural-metal preparers and related trades workers
722	Blacksmiths, tool-makers and related trades workers
723	Machinery mechanics and fitters
724	Electrical and electronic equipment mechanics and fitters
731	Precision workers in metal and related materials
734	Craft printing and related trades workers
741	Food processing and related trades workers
812	Metal-processing-plant operators
814	Wood-processing- and paper-making-plant operators
815	Chemical-processing-plant operators
816	Power-production and related plant operators
821	Metal- and mineral-products machine operators
822	Chemical-products machine operators
823	Rubber- and plastic-products machine operators
824	Wood-products machine operators
825	Printing- binding- and paper-products machine operators
826	Textile-, fur-, and leather-products machine operators
827	Food and related products machine operators
828	Assemblers
829	Other machine operators and assemblers
831	Locomotive-engine drivers and related workers
832	Motor-vehicle drivers
833	Agricultural and other mobile-plant operators
912	Helpers and cleaners
913	Helpers in restaurants
914	Doorkeepers, newspaper and package deliverers and related workers
915	Garbage collectors and related workers
919	Other sales and services elementary occupations
932	Manufacturing labourers
933	Transport labourers and freight handlers