# Measuring Family (Dis)Advantage: Lessons from Detailed Parental Information

Sander de Vries*

June 14, 2025

**Abstract**

This paper provides new insights on the importance of family background by linking income, education, and crime of 1.7 million Dutch children to exceptionally rich information about their parents and aunts and uncles. I provide a detailed analysis of how these outcomes vary across family backgrounds and show that incorporating all information explains substantially more variation in income (+56%), education (+102%), and violent crime (+161%) than analyses using parental income only. Neighborhoods and migration background explain a limited portion of the income disparities. Instead, a comparison with adoptees highlights the role of pre-birth factors in driving intergenerational income dependence.

**Keywords:** intergenerational mobility, inequality of opportunity

**JEL Codes:** I24, J24, J62

*Department of Economics, Vrije Universiteit Amsterdam, s.de.vries@vu.nl. I gratefully acknowledge valuable comments from Nadine Ketel, Maarten Lindeboom, Erik Plug, Paul Hufe, Gustave Kenedi, and conference and seminar participants in Amsterdam, Utrecht, The Hague, Tokyo, London, Canazei, Colchester, and Antwerp. The non-public micro data used in this paper are available via remote access to the Microdata services of Statistics Netherlands (project agreement 8674).

# 1 Introduction

Understanding how income is distributed across different types of families is crucial for evaluating inequality and informing policy. Disparities by family background are widely perceived as unfair and preferences for redistribution and equality-of-opportunity policies depend on their believed extent (Alesina et al., 2018). Moreover, identifying the family types where children consistently attain high or low incomes can help policymakers better target interventions to reduce such disparities. Measuring family (dis)advantage can thus provide important insights into both the need for such policies and their effective targeting.

Yet, the most commonly used measures of intergenerational mobility provide limited insights on how income is distributed across family backgrounds. Estimates based on parent-child income associations capture just one dimension of family background and may understate its broader influence. To address this limitation, economists also frequently estimate sibling correlations (Solon (1999)). However, these correlations capture all shared influences between siblings, including factors unrelated to parents, and may greatly overstate the role of the family (Collado et al. (2023)). Most importantly, both approaches are silent on where in the income distribution family background matters most, or what differentiates families whose children succeed economically from those who do not.

This paper aims to fill this gap by leveraging administrative data with exceptionally rich family background information. Specifically, I link the long-run incomes of over 1.7 million Dutch children to detailed information on their fathers, mothers, and aunts and uncles, including income, assets, debt, occupation, education, criminal records, health, and family structure. Although this extensive dataset does not capture all potentially relevant dimensions of family background, it greatly expands upon the scope of previous studies. I use all family information in a flexible machine learning model to predict child income ranks. This predictive model provides a non-parametric description of how children's expected income ranks vary by their family background characteristics.

I show that incorporating all family information reveals substantially greater disparities

between families than analyses based on parental income alone. While parental income explains 10.5 percent of the variation in child income ranks, the full set of family characteristics accounts for 16.4 percent, a 56 percent increase in explanatory power.[1] To illustrate, the 0.5 percent of children with the lowest expected incomes based on parental income alone have an average observed income rank of 31; under the comprehensive model, this drops to 19. The increase in explanatory power is even larger for predicting children's education (102%) and crime (161%). These findings highlight the value of a multidimensional approach for accurately quantifying family-driven inequalities in important child outcomes.

Another key contribution of this paper is to provide granular insights on the distribution of expected income and the corresponding family characteristics. The income and wealth of parents and aunts or uncles are the strongest income predictors, although other variables also make meaningful contributions. Children with the highest (lowest) expected incomes face a combination of (dis)advantages. For instance, the 0.5 percent of children with the highest expected incomes have an observed mean income rank of 77 and come from families with favorable characteristics across nearly all dimensions, with particularly high parental income and wealth. In contrast, children in the bottom 0.5 percent (mean rank 19) have parents who are often young, separated, and who have low income and wealth, limited education, poor health, and criminal records, with similar disadvantages among aunts and uncles.

Finally, I consider some broad channels that could drive these disparities. I find that predicted income differences remain accurate among children from the same neighborhood or migrant group, suggesting that these factors cannot explain the disparities. Next, I compare the incomes of international adoptees raised in families with different levels of advantage. The results indicate that being raised from infancy in a family associated with a 1 rank higher expected income for own-birth children increases the income of adoptees by only 0.3 ranks. This provides strong evidence that a substantial share of family (dis)advantage is rooted in pre-birth factors.

[1] This analysis also reveals that the rank-rank correlation in the Netherlands (0.32) is considerably higher than previously thought (0.22, Van Elk et al. (2024)).

This paper contributes to the growing strand of literature measuring the importance of family background with multiple family characteristics, which encompasses papers from both the intergenerational mobility and inequality of opportunity literature.[2] Most closely related are recent papers using machine learning to predict children's income (Blundell and Risa (2019), Brunori et al. (2023b), Brunori et al. (2024), Chang et al. (2025)).[3] I follow a similar approach, but consider far more detailed information than previous papers, including the value of specific types of assets and debt, detailed occupation information, health, criminal behavior, family structure, and extended family outcomes, all of which are shown to be significant predictors.[4] Moreover, while previous studies focus on aggregated intergenerational mobility or inequality of opportunity estimates, this paper provides a substantially more granular analysis by reporting the full distribution of expected incomes alongside the corresponding family background characteristics. These results offer a more nuanced understanding of the magnitude of unfair inequality and can be used by policymakers to more effectively target policy to reduce such inequalities.

# 2 Data

*Sample.* I use administrative data from Statistics Netherlands covering the full Dutch population.[5] The main sample consists of all children born in the Netherlands between 1980 and 1989, excluding 3.4 percent with missing income data, resulting in 1,703,038 observations. For the education and crime analyses, I focus on birth cohorts 1985–1989 due to unsuitable

[2]For comprehensive reviews, see Black and Devereux (2011) and Mogstad and Torsvik (2023) on intergenerational mobility, and Roemer and Trannoy (2016) and Ramos and Van de Gaer (2016) on inequality of opportunity.

[3]Other related contributions are Vosters and Nybom (2017) and Vosters (2018), who aggregate information from multiple measures into a least-attenuated linear estimator of persistence in a latent variable framework, Adermon et al. (2021), who propose a new estimator of intergenerational mobility based on outcomes from multiple family members, and Eshaghnia et al. (2022), who measure mobility using expected lifetime income, which is based on multiple parental characteristics.

[4]This paper also stands out by including detailed information about mothers. Remarkably few studies use information about fathers or mothers separately. Other recent works addressing this gap are Brandén et al. (Forthcoming), Ahrsjö et al. (2023), and Althoff et al. (2024).

[5]Access is granted through a secure remote facility under a confidentiality agreement.

education and crime records for earlier periods. After excluding 0.5 percent with missing education records, the education sample includes 908,876 children. The crime analysis focuses on boys only, yielding 463,625 observations.

*Children's household income.* The main outcome in this paper is a child's long-run gross household income rank. I focus on household income because it provides a reliable measure of economic resources even in the case of non-participation in the labor market and it is commonly used in other intergenerational mobility studies (Chadwick and Solon (2002)). Household incomes are observed between 2003 and 2023 and encompass all income from employment, entrepreneurship, and capital, as well as income insurance payments, social security benefits, conditional transfers, receiving income transfers, and employers' and employees' contributions to social insurance premiums.[6] I measure income in 2024 euros, adjusting for inflation using the consumer price index.

I construct a proxy for children's lifetime household income by averaging their household income from age 30 onward.[7] This approach reduces measurement error from transitory income shocks (Mazumder, 2005) and life-cycle bias (Haider and Solon (2006), Nybom and Stuhler (2017)). I observe income up to age 43 for the oldest cohort (born in 1980) and up to age 34 for the youngest cohort (born in 1989). On average, children have nine income observations, with 96 percent having at least five. I then rank children within birth-years based on their lifetime household income. I also present results for various alternative measures to evaluate the sensitivity of the results due to these choices.

*Children's education and crime.* I use individuals' highest completed education to construct a years-of-education variable according to the conversion table in Appendix E.

[6]Some children still live with their parents when I measure their income. In these cases, I define the income of the children as their gross personal income and that of the parents as the household income minus the total gross personal income of the children who still live at home.

[7]I exclude years with yearly household income below €1,000 (0.6%), as these cases typically correspond to wealthy entrepreneurs with business losses.

The crime register data contains all offenses reported to the police between 2005 and 2022, including the individual identifier of the suspected offender(s). The crime outcome is an indicator of whether a child has been suspected of any *violent* crime at ages 20 to 33. This is the longest age window for which I can accurately observe children's criminal behavior and corresponds to prime ages when children commit crimes. I focus on violent crimes because of their high societal costs.

*Parental household income.* I estimate each parent's lifetime household income by averaging their annual incomes up to age 60. Since most parents were born in the 1950s, their first incomes are typically observed around their late 40s. On average, fathers have 12 income observations and mothers 14. Following Chetty et al. (2014), parental income is defined as the average of the father's and mother's lifetime household income. If only one parent's income is observed, I use that parent's income. The parental income rank is based on the position within the parental income distribution of all children in the analysis sample.

*Other explanatory variables.* Table 1 describes how the other variables are classified into seven categories. Except for household income and wealth, which are measured at the household level, all variables are included for the father and the mother separately. Altogether, the set comprises 75 continuous variables, 8 binary indicators, and 2 categorical variables each containing 68 distinct categories. Appendix B provides descriptive statistics for the core sample, including all explanatory variables, as well as a detailed explanation of how the explanatory variables are constructed.

Although the data are rich, they come with two limitations. First, some parental outcomes are observed only after their children have left the household. Consequently, my results may underestimate the importance of family background compared to a model that includes information on parents' resources and well-being during their children's formative years. Nonetheless, many parental characteristics are highly persistent over the life cycle,

making them a reasonable proxy for the family environment at earlier ages.[8]

Second, despite the extensive coverage of variables, some missing values persist. Most importantly, education records for the parents' generation are incomplete, as systematic recording began in the 1980s. Extended family outcomes are also unavailable for some children, often because their parents have no siblings or their grandparents cannot be identified, making it impossible to link to aunts or uncles. To preserve the full sample, I use indicators to denote missing information instead of excluding incomplete observations.

# 3 Model training and evaluation

The objective is to train a predictive model that accurately predicts the Conditional Expectation Function (CEF) of children's incomes given the family background characteristics. A key challenge is that the true functional form of the CEF is unknown. Variables may enter in a non-linear manner or interact with other variables. In these cases, flexible machine learning methods can outperform linear regression models. Accordingly, I employ gradient-boosted decision trees to generate these predictions (Friedman (2001)). Tree-based methods offer the additional advantage of providing Shapley value–based measures of variable importance even with a large number of predictors, which is infeasible with most alternative approaches (Lundberg and Lee (2017), Lundberg et al. (2020)). In Appendix D, I provide a simple framework that relates the approach in this paper to intergenerational mobility estimates, sibling correlations, and inequality of opportunity estimates.

For each analysis requiring a separate predictive model, I randomly split the sample into a training set (80 percent) and a test set (20 percent). I perform 5-fold cross-validation on the training data to determine the optimal parameter values, and then train a final model on the full training set using these parameters. This model is then applied to observations from the test data to estimate the out-of-sample explanatory power ($R^2$). Generally, all results in

[8]This is supported by Eshaghnia et al. (Forthcoming), who show that differences in intergenerational mobility estimates due to different types of resources being analyzed are much larger than differences due to the age of the children at which these resources are measured.

this paper that rely on predictions are based on observations from the test data.

# 4 Main Results

## 4.1 Intergenerational Income Mobility in the Netherlands

As a baseline analysis, I first estimate the relationship between child and parent income. A regression of child income rank on parental income rank yields a slope coefficient of 0.32 ($R^2$ = 10.5, see Table B1). This estimate positions the Netherlands among the OECD countries with relatively strong persistence. Intergenerational persistence in the Netherlands is higher than in Sweden, Denmark, Australia, Norway, Germany, and Canada (0.20-0.24), similar to France, Italy, and the UK ($\sim$ 0.30), and lower than in the US (0.36).[9] Despite the Netherlands' relatively low-income inequality and accessible education, intergenerational persistence appears surprisingly high.[10]

A detailed plot of child income ranks versus parental income rank and alternative mobility estimates are presented in Appendix B. These include the commonly used intergenerational income elasticity (IGE), which also equals 0.32, and separate analyses for sons and daughters using personal income ranks, which both yield estimates of 0.29. Moreover, I vary the number of years over which income is measured and the specific periods in parents' or children's lives when their incomes are recorded. These robustness checks suggest that the estimate is robust to measurement error and lifecycle bias. This also implies that any additional explanatory power gained from incorporating more variables is unlikely due to these variables merely correcting for measurement error in the parental lifetime income.

[9]See (in the same order): Heidrich (2017), Helsø (2021), Deutscher and Mazumder (2020), Bratberg et al. (2017), Corak (2020), Kenedi and Sirugue (2023), Acciari et al. (2022), Rohenkohl (2023), Davis and Mazumder (2024).

[10]The estimated rank-rank correlation is higher than the 0.22 estimate reported in Van Elk et al. (2024). They use fewer years of income information for parents and children (3 years), measure income at younger ages for children, do not trim incomes below €1000, and apply stricter sample selection restrictions, all of which may result in smaller estimates.

## 4.2 Including Detailed Parental Information

The previous section analyzes how children's incomes vary by parental income. This section shows how children's incomes vary by their broader family background characteristics detailed in Section 2.

To quantify the increase in family-driven inequality when adding the broader family background information, I compare the explanatory power of a model using only parental income with that of a model incorporating all explanatory variables. Both models are trained and evaluated on the same training and test data. For the income-only model, I non-parametrically predict a child's income rank in the test data by the mean income rank of all children in the training data with the same parental income rank and year of birth. Like the linear regression in the previous section, this model achieves an explanatory power of 10.5 percent. The predictions using all explanatory variables are generated by a tuned gradient-boosted decision tree, as described in Section 3.

Adding all information about the parents reveals substantially stronger intergenerational dependence. The comprehensive model achieves an explanatory power of 16.4 percent, marking a 56 percent increase compared to the income-only model. To put this into perspective, an increase in the rank-rank correlation from 0.32 to 0.40 would result in the same increase in $R^2$.[11] While this may seem modest, it is considerable, considering the difference in rank-rank correlation between Sweden (high mobility) and the US (low mobility) is about 0.16. Moreover, the increase in $R^2$ far exceeds the gain achieved from reducing attenuation bias in an income rank-rank regression.[12] This source of measurement error has received considerable attention in the literature (Mazumder (2005), Nybom and Stuhler (2017)).When the goal is to quantify income disparities between families, adding more information about parents is thus more valuable than constructing a more accurate proxy of lifetime income.

Figure 1 provides a binscatter plot of children's mean income ranks, sorted from lowest to

[11] I use here that in a rank-rank regression, $R^2 = \beta^2$ (i.e. $0.405^2 - 0.324^2 = 0.164 - 0.105 = 0.059$).

[12] Table B2 columns 1 and 9 shows that using all available income data versus one year of income data in a rank-rank regression increases the $R^2$ from 7.6% to 10.5%.

highest predicted income. The X-axis divides the test dataset into 200 bins, each containing approximately 1,700 children, based on their predicted income ranks within their cohort. The Y-axis reports the average observed income rank for each bin. The blue dots represent children grouped by predicted income using parental income alone, while the orange dots reflect groupings based on predictions from the comprehensive model.

The comprehensive model identifies considerably greater income disparities by family background. For instance, in the income-only model, the 1 percent of children with the lowest expected income have an average income rank of 31. With the comprehensive model, this drops to 19. Similarly, for the top 1 percent, the income-only model estimates an average rank of 70, while incorporating additional family background information raises this to 75.

The results also highlight that a small group of children are particularly advantaged or disadvantaged. For example, even within the top 1 percent, there are striking differences: those in the top 0.5 percent of expected incomes reach an average rank of 77.1, which is 3.5 ranks higher than the next 0.5 percent. Similarly, the bottom 0.5 percent of expected incomes have an average rank of 18.8, which is 2.3 ranks lower than the next 0.5 percent. To put this in perspective, in analyses with parental income ranks only, such disparities would imply local rank-rank correlations far above 1.[13]

## 4.3 What Predicts Child Income?

Table 2 shows how family characteristics vary across the expected income distribution, focusing on the most advantaged and disadvantaged children.[14] The first and last four columns include the 10 percent of children with the lowest and highest expected incomes, while the fifth column contains all children in between. Row 1 shows the corresponding mean income ranks and the remaining rows report the average family background characteristics.

Table 2 shows that children at the extremes face multiple (dis)advantages. The first four

[13]To the best of my knowledge, the highest local intergenerational mobility coefficient is documented by Björklund et al. (2012), which is 0.9 among the top 0.1% of parental incomes in Sweden.

[14]As it is not feasible to report descriptive statistics for *all* included explanatory variables, I selected these variables to broadly cover all different dimensions of family background from Table 1.

columns show that the most disadvantaged children have parents with low income and wealth and who are often young, separated, minimally educated, have high health expenditures, and are often suspected of crimes. Their aunts and uncles also have low income and wealth. In contrast, the most advantaged children have family members with favorable characteristics across all these dimensions.

Although the variables in Table 2 are all correlated with child income, they are not equally good predictors. Appendix C presents a detailed graph illustrating the variable importance of the 30 most predictive variables, calculated using Shapley values. This analysis reveals two insights. First, all variables except for whether the father or mother is identified contribute to the predictions, indicating that each adds valuable information to the analysis. Second, income and wealth variables for parents and extended family exert the strongest influence on predictions. The top nine predictors fall into these categories, underscoring the essential role of income and wealth data in measuring disparities by family background.

## 4.4 Relation To Other Methods

*Sibling correlation.* The correlation between two randomly drawn siblings provides an upper bound on the explanatory power of any predictive model that solely includes variables that are equal between siblings (as is the case in this paper).[15] The sibling correlation in income in the core sample is 0.31. This implies that the included family information explains about 50 percent of the sibling correlation (0.16/0.31). The remaining half may be explained by other factors shared between siblings, such as community influences, shocks, or sibling spillovers, that are uncorrelated with the family background variables.

Sections 4.2 and 4.3 highlight two advantages of the prediction approach over sibling correlations. First, the predictions only rely on observable family characteristics, whose contributions can be quantified. Instead, the sibling correlation relies on many unobservable factors shared between siblings, which complicates its interpretation. Second, the prediction

[15]This follows from the fact that the sibling correlation equals the (adjusted) $R^2$ of a regression of income on family fixed effects. See Appendix D for a more detailed discussion.

approach produces a full distribution of expected incomes, allowing for detailed analysis of intergenerational dependence. The sibling correlation approach relies on many imprecisely estimated family fixed effects, making it unsuitable for such a distributional analysis.

*Inequality of opportunity.* Figure 1 presents the distribution of expected incomes without imposing any normative judgment on its fairness. The inequality of opportunity literature goes a step further by mapping this distribution and that of observed incomes into summary measures that reflect distinct normative perspectives on their fairness (Brunori et al. (2023a), Brunori et al. (2024)). These measures rely on the choice of inequality index, which determines how disparities at different points of the (predicted) income distribution are weighted. Below, I present inequality of opportunity estimates using several of these indices.

Because the scale of the income differences play a central role in normative evaluations, this literature uses income levels rather than ranks. Figure A1 presents results analogous to those in Figure 1, but with models predicting income levels. Using these predictions, I find that inequality of opportunity measured by the Gini index is 40 percent, whereas it is notably lower using Theil's index, the Mean Logarithmic Deviation index, or the variance (11 to 15 percent).[16] The Gini estimate is substantially higher because it places more weight on disparities in the middle of the distribution, where most predicted incomes are concentrated. This highlights that normative assessments of the income differences by family background depend crucially on how one weights these disparities across the distribution.

## 4.5 Gender Differences and Robustness

*Gender differences.* Figures A2 and A3 present results from predictive models trained to predict sons' and daughters' income ranks separately. The explanatory power for predicting household income ranks is similar between genders, and for predicting personal income ranks,

[16]These estimates are computed by applying each inequality index to (a) the observed income distribution and (b) the distribution of predicted incomes in the test set. Ex-ante inequality of opportunity estimates equal the ratio of (b) to (a) (Brunori et al. (2023b)).

it is slightly higher for daughters. The household income predictions for boys and girls from the same family are almost perfectly correlated (97%). This suggests that not only the level of intergenerational dependence, but also the family background characteristics contributing to it, are similar between genders.

*Robustness.* Machine learning models require sufficiently large samples to effectively tune their parameters and inadequate sample size can result in suboptimal performance. Table A2 shows that explanatory power declines with smaller samples but stabilizes once at least 40 percent of the data are used. This suggests that the estimates reported are unlikely to suffer from downward bias due to insufficient sample size.

Table A3 varies the number of years and ages at which child income is measured. The results from this analysis mimic that of the robustness of the rank-rank correlation in Appendix B. Explanatory power attenuates when fewer years of income are used, but stabilizes once at least five years of income are used. It also decreases somewhat when income is measured exclusively in the early 30s, but when I re-estimate the model using only incomes beyond age 32, then the overall estimate is virtually identical to the main results specification. This indicates that the influence of attenuation or life-cycle bias is likely minimal.

Finally, I assess the importance of the missing education records. I first train the model on the subset of children whose parents' education is observed ($n = 1,093,245$, $R^2 = 17.28$). I then re-train the model on the same sample after removing all education variables for both parents and extended family ($R^2 = 17.13$). The resulting drop in $R^2$ is only 0.15 percentage points, indicating that the remaining variables already capture most of the educational variation across families. This suggests that the explanatory power of the model would increase only marginally if complete education data were available.

## 4.6 Predicting Education and Crime

To examine whether the value of additional family information varies across outcomes, this section presents results for children's completed education and violent criminal behavior. As

violent crimes are predominantly committed by men (85 percent), I focus on men's criminal behavior only.

Figure 2 (a) presents the results for education. The explanatory power of the income-only model is 12.6 percent.[17] Incorporating all explanatory variables significantly boosts explanatory power for education, doubling it to 25.5 percent. This increase in explanatory power is considerably larger than for predicting income.

The comprehensive model reveals large disparities in children's education by family background. For example, children with the 5 percent lowest predicted education levels have on average less than 12 years of education, frequently dropping out without qualifications, whereas children with the 5 percent highest predicted education levels have on average 17.1 years of education, corresponding to an undergraduate degree.

Figure 2(b) shows a similarly large increase in explanatory power for crime, from 3.9 percent for the model that incorporates income only to 10.2 percent for the comprehensive model. The results indicate that violent crime is highly concentrated in certain families. For a small subset of children, the probability of being suspected even exceeds fifty percent. A simple calculation shows that the 20 percent of boys with the highest crime risk in Figure 2 (b) account for 50 percent of all boys who have been suspected of a violent crime between the ages of 20 and 33.

Overall, the findings above imply that a multidimensional approach is even more valuable for measuring the importance of family background for education and crime than for income.

[17]Intergenerational mobility studies often apply regressions of child education on the highest education of the parents. Applying this regression to a subsample of children for whom at least one parent's education is observed, I find an explanatory power of 11.7 percent.

# 5 Drivers of Family (Dis)Advantage

## 5.1 Neighborhoods and Migration Background

Family background is closely linked to both the neighborhoods in which children grow up and their migration background. This section assesses the extent to which these factors can explain the disparities shown in Figure 1.

Table 3 column 2 shows a regression of child income on predicted income and neighborhood fixed effects, corresponding to the neighborhood where children were registered at age 15.[18] Without these fixed effects, the coefficient is equal to 1 by construction (column 1). The neighborhood fixed effects ensure that only children from the same neighborhood are compared, which may result in a decrease in the coefficient if the income predictions capture differences between neighborhoods. However, the coefficient falls only slightly to 0.94, implying that a 1-rank increase in predicted income is associated with a 0.94-rank increase in actual income even *within* neighborhoods. This shows that income disparities by family background are almost equally strong within neighborhoods as between neighborhoods.[19]

Second- and third-generation migrants comprise 21 percent of the sample and have, on average, lower incomes than natives. Consequently, part of the disparities in Figure 1 could reflect differences in migration background correlated with the family background variables. To test this, column 2 includes fixed effects for the region of origin of the father, mother, and grandparents (when available), restricting the comparisons to individuals whose family members share the same origin.[20] The coefficient in column 2 is nearly one, indicating that income disparities by family background are just as pronounced within migrant groups as across them. Column 3, which restricts the analysis to second- and third-generation migrants

[18]The neighborhood code is based on the most granular level of Statistics Netherlands' neighborhood classifications. The mean and median neighborhood sizes are 1500 and 900 individuals, respectively.

[19]These results are consistent with papers that find that neighborhoods can explain only a limited fraction of the sibling correlation (Solon et al. (2000), Page and Solon (2003), Raaum et al. (2006), Bingley et al. (2021)).

[20]I distinguish eight regions: the Netherlands, Morocco, Turkey, Surinam, Dutch Antilles, Western Europe, Eastern Europe, and others.

only, yields a similarly high estimate.

Taken together, neighborhoods and migration background can explain only a small portion of the disparities by family background. The most likely hypothesis is therefore that parental inputs are the main drive behind the disparities observed in the main results.

## 5.2 The Post-birth Environment

Parental inputs include post-birth factors such as general care and household resources, but also pre-birth factors like genetic endowments and the prenatal environment. To understand the relative importance of these two inputs in driving the disparities observed in Figure 1, this section examines the causal effect of being assigned shortly after birth to a family associated with a 1 rank higher predicted income.

To answer this question, I use a sample of 4,935 international adoptees born between 1980 and 1989 and who arrived in the Netherlands within six months of birth.[21] These children are not genetically related to their adoptive parents and were not cared for by them during pregnancy and shortly after birth, but have been raised by them since they were at most 6 months old. This unique context makes them an interesting group for studying the importance of the post-birth environment.[22]

The primary assumption is that adoptees were effectively randomly assigned to parents. Although limited institutional information on matching procedures from this period restricts a comprehensive assessment of this assumption, two considerations support its plausibility. First, the excess demand for infant adoptees in the 1980s likely discouraged selective placement, as prioritizing specific characteristics would have significantly increased already long

[21] Although the Netherlands lacks an adoption register, Statistics Netherlands developed a reliable method to identify adoptees. They sent a survey to a random subset of all plausible adoptees to verify their method. Overall, 97.7 percent of respondents in my sample confirmed they were adopted ($n = 778$). Nevertheless, a minor fraction of plausible adoptees may not be adopted. This may induce a small upward bias in the estimate in Column 4 of Table 3.

[22] The approach here is commonly used in previous papers (e.g. Sacerdote (2011), Holmlund et al. (2011), Fagereng et al. (2021)). This section extends previous results by characterizing family background with a multidimensional measure - the income prediction, which is based on many family background variables - and by focusing on children's long-run income. Despite its central role in intergenerational mobility analyses, this dimension has been overlooked in studies using international adoptees.

waiting times.[23] Second, I report estimates from various specifications in Table A4 that include controls for gender, age at migration, and fully interacted fixed effects for the country and year of adoption, which are all observable characteristics of the child at the time of adoption. These estimates are effectively unchanged, indicating that selective placements based on these observable characteristics are of limited empirical importance.

Column 4 of Table 3 shows that being raised in a family that is associated with a 1 rank higher income for own-birth children increases the income rank of adoptees by only 0.27. Assuming no selection bias and generalizability towards the broader population, this estimate would suggest that around 27% of the disparities in Figure 1 are shaped by the post-birth environment.

Although adoptees and their adoptive parents are clearly not representative of the broader population, I also offer three reasons why external validity concerns may not be overly severe. First, column 5 in Table 3 shows that the association between realized and predicted income stays close to one for own-birth children in families with at least one adopted child. This indicates that differences in the predictability of income between adoptees and own-birth children are not driven by fundamental differences between families with and without adoptees. Second, while there are no highly disadvantaged adoptive families, Table A5 reveals substantial variation in the characteristics of adoptive families, spanning a broad range of the general population. Third, as shown in Table 3, the predictive model performs well for children with a migration background, indicating that the non-native status of adoptees does not substantially threaten generalizability.

# 6 Conclusion

This paper provides a detailed analysis of how children's key economic outcomes vary across family backgrounds. I show that including family background characteristics beyond parental

[23] Waiting times during this period could span several years. See, for example, Rapport Commissie Onderzoek Interlandelijke Adoptie (in Dutch, 2021), `https://www.rijksoverheid.nl/documenten/rapporten/2021/02/08/tk-bijlage-coia-rapport`.

income considerably increases estimates of intergenerational dependence. This increase stems from better identification of highly (dis)advantaged children, whose families exhibit (un)favorable outcomes across multiple dimensions. Neighborhoods and migration background do not explain the observed disparities between families. Instead, pre-birth factors appear to be an important driver.

The multidimensional approach should be seen as a complement to, rather than a substitute of, simpler measures of intergenerational mobility. Simpler measures are easier to interpret and compare, and prior work shows that they yield similar regional rankings even if they understate differences in levels (Blundell and Risa (2019), Deutscher and Mazumder (2023), Adermon et al. (2025)). However, when the aim is to obtain a more precise view of the *level* of (dis)advantage experienced by different children and what characterizes their families, this paper demonstrates the value of a richer, multidimensional approach.

# References


**Acciari, Paolo, Alberto Polo, and Giovanni L. Violante.** 2022. "And Yet It Moves: Intergenerational Mobility in Italy." *American Economic Journal: Applied Economics* 14 (3): 118–163.

**Adermon, Adrian, Gunnar Brandén, and Martin Nybom.** 2025. "The Relationship between Intergenerational Mobility and Equality of Opportunity." IFAU Working Paper No. 2025:2.

**Adermon, Adrian, Mikael Lindahl, and Mårten Palme.** 2021. "Dynastic Human Capital, Inequality, and Intergenerational Mobility." *American Economic Review* 111 (5): 1523–1548.

**Ahrsjö, Ulrika, René Karadakic, and Joachim Kahr Rasmussen.** 2023. "Intergenerational Mobility Trends and the Changing Role of Female Labor." arXiv preprint, arXiv:2302.14440.

**Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso.** 2018. "Intergenerational Mobility and Preferences for Redistribution." *American Economic Review* 108 (2): 521–554.

**Althoff, Lukas, Harriet Brookes Gray, and Hugo Reichardt.** 2024. "The Missing Link(s): Women and Intergenerational Mobility." Stanford University Working Paper, `https://lukasalthoff.github.io/pdf/igm_mothers.pdf`.

**Angrist, Joshua D., and Jörn-Steffen Pischke.** 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

**Bingley, Paul, and Lorenzo Cappellari.** 2019. "Correlation of Brothers' Earnings and

Intergenerational Transmission." *The Review of Economics and Statistics* 101 (2): 370–383.

**Bingley, Paul, Lorenzo Cappellari, and Konstantinos Tatsiramos.** 2021. "Family, Community and Long-Term Socio-Economic Inequality: Evidence from Siblings and Youth Peers." *The Economic Journal* 131 (636): 1515–1554.

**Björklund, Anders, Jesper Roine, and Daniel Waldenström.** 2012. "Intergenerational Top Income Mobility in Sweden: Capitalist Dynasties in the Land of Equal Opportunity?" *Journal of Public Economics* 96 (5): 474–484.

**Black, Sandra E., and Paul J. Devereux.** 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics*, edited by Card, David, and Orley Ashenfelter Volume 4. 1487–1541.

**Blundell, Jack, and Erling Risa.** 2019. "Income and Family Background: Are We Using the Right Models?" Working Paper, available at SSRN: `https://ssrn.com/abstract=3269576`.

**Brandén, Gunnar, Martin Nybom, and Kelly Vosters.** Forthcoming. "Like Mother, Like Child? The Rise of Women's Intergenerational Income Persistence in Sweden and the United States." *Journal of Labor Economics*.

**Bratberg, Espen, Jonathan Davis, Bhashkar Mazumder, Martin Nybom, Daniel D. Schnitzlein, and Kjell Vaage.** 2017. "A Comparison of Intergenerational Mobility Curves in Germany, Norway, Sweden, and the US." *The Scandinavian Journal of Economics* 119 (1): 72–101.

**Brunori, Paolo, Francisco H.G. Ferreira, Guido Neidhöfer, and UNU-WIDER.** 2023a. "Inequality of Opportunity and Intergenerational Persistence in Latin America." WIDER Working Paper 2023.

**Brunori, Paolo, Francisco H.G. Ferreira, and Pedro Salas-Rojo.** 2024. "Inherited Inequality: A General Framework and a 'Beyond-Averages' Application to South Africa." IZA Discussion Paper No. 17203.

**Brunori, Paolo, Paul Hufe, and Daniel Mahler.** 2023b. "The Roots of Inequality: Estimating Inequality of Opportunity from Regression Trees and Forests." *The Scandinavian Journal of Economics* 125 (4): 900–932.

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

**Chang, Yoosoon, Steven N. Durlauf, Bo Hu, and Joon Park.** 2025. "Accounting for Individual-Specific Heterogeneity in Intergenerational Income Mobility." NBER Working Paper 33349.

**Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez.** 2014. "Where Is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *The Quarterly Journal of Economics* 129 (4): 1553–1623.

**Collado, M Dolores, Ignacio Ortuño-Ortín, and Jan Stuhler.** 2023. "Estimating Intergenerational and Assortative Processes in Extended Family Data." *The Review of Economic Studies* 90 (3): 1195–1227.

**Corak, Miles.** 2020. "The Canadian Geography of Intergenerational Income Mobility." *The Economic Journal* 130 (631): 2134–2174.

**Davis, Jonathan MV, and Bhashkar Mazumder.** 2024. "The Decline in Intergenerational Mobility after 1980." *Review of Economics and Statistics* 1–47.

**Deutscher, Nathan, and Bhashkar Mazumder.** 2020. "Intergenerational Mobility across Australia and the Stability of Regional Estimates." *Labour Economics* 66 101861.

**Deutscher, Nathan, and Bhashkar Mazumder.** 2023. "Measuring Intergenerational Income Mobility: A Synthesis of Approaches." *Journal of Economic Literature* 61 (3): 988–1036.

**Eshaghnia, Sadegh, James J. Heckman, and Rasmus Landersø.** Forthcoming. "The Impact of the Level and Timing of Parental Resources on Child Development and Intergenerational Mobility." *Journal of Labor Economics*.

**Eshaghnia, Sadegh, James J. Heckman, Rasmus Landersø, and Rafeh Qureshi.** 2022. "Intergenerational Transmission of Family Influence." NBER Working Paper 30412.

**Fagereng, Andreas, Magne Mogstad, and Marte Rønning.** 2021. "Why Do Wealthy Parents Have Wealthy Children?" *Journal of Political Economy* 129 (3): 703–756.

**Friedman, Jerome H.** 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics* 29 (5): 1189–1232.

**Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan.** 2022. "Measuring the Completeness of Economic Models." *Journal of Political Economy* 130 (4): 956–990.

**Haider, Steven, and Gary Solon.** 2006. "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review* 96 (4): 1308–1320.

**Heidrich, Stefanie.** 2017. "Intergenerational Mobility in Sweden: A Regional Perspective." *Journal of Population Economics* 30 (4): 1241–1280.

**Helsø, Anne-Line.** 2021. "Intergenerational Income Mobility in Denmark and the United States*." *The Scandinavian Journal of Economics* 123 (2): 508–531.

**Holmlund, Helena, Mikael Lindahl, and Erik Plug.** 2011. "The Causal Effect of Parents' Schooling on Children's Schooling: A Comparison of Estimation Methods." *Journal of Economic Literature* 49 (3): 615–651.

**Kenedi, Gustave, and Louis Sirugue.** 2023. "Intergenerational Income Mobility in France: A Comparative and Geographic Analysis." *Journal of Public Economics* 226 104974.

**Lundberg, Scott M., Gabriel Erion, Hugh Chen et al.** 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature machine intelligence* 2 (1): 56–67.

**Lundberg, Scott M, and Su-In Lee.** 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, Volume 30.

**Mazumder, Bhashkar.** 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *The Review of Economics and Statistics* 87 (2): 235–255.

**Mogstad, Magne, and Gaute Torsvik.** 2023. "Family Background, Neighborhoods, and Intergenerational Mobility." In *Handbook of the Economics of the Family*, edited by Lundberg, Shelly, and Alessandra Voena Volume 1. 327–387.

**Nybom, Martin, and Jan Stuhler.** 2017. "Biases in Standard Measures of Intergenerational Income Dependence." *The Journal of Human Resources* 52 (3): 800–825.

**Page, Marianne E., and Gary Solon.** 2003. "Correlations between Sisters and Neighbouring Girls in Their Subsequent Income as Adults." *Journal of Applied Econometrics* 18 (5): 545–562.

**Raaum, Oddbjørn, Kjell G. Salvanes, and Erik Ø. Sørensen.** 2006. "The Neighbourhood Is Not What It Used to Be." *The Economic Journal* 116 (508): 200–222.

**Ramos, Xavier, and Dirk Van de Gaer.** 2016. "Approaches to Inequality of Opportunity: Principles, Measures and Evidence." *Journal of Economic Surveys* 30 (5): 855–883.

**Roemer, John E., and Alain Trannoy.** 2016. "Equality of Opportunity: Theory and Measurement." *Journal of Economic Literature* 54 (4): 1288–1332.

**Rohenkohl, Bertha.** 2023. "Intergenerational Income Mobility: New Evidence from the UK." *The Journal of Economic Inequality* 21 (4): 789–814.

**Sacerdote, Bruce.** 2011. "Chapter 4 - 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*, edited by Hanushek, Eric A., Stephen Machin, and Ludger Woessmann Volume 3. 249–277.

**Shapley, L. S.** 1953. "Stochastic Games." *Proceedings of the National Academy of Sciences* 39 (10): 1095–1100.

**Solon, Gary.** 1999. "Chapter 29 - Intergenerational Mobility in the Labor Market." In *Handbook of Labor Economics*, edited by Ashenfelter, Orley C., and David Card Volume 3. 1761–1800.

**Solon, Gary, Marianne E. Page, and Greg J. Duncan.** 2000. "Correlations between Neighboring Children in Their Subsequent Educational Attainment." *The Review of Economics and Statistics* 82 (3): 383–392.

**Van Elk, Roel Adriaan, Egbert Jongen, Patrick Koot, and Alice Zulkarnain.** 2024. "Intergenerational Mobility of Immigrants in the Netherlands." IZA Discussion Paper No. 17035.

**Vosters, Kelly.** 2018. "Is the Simple Law of Mobility Really a Law? Testing Clark's Hypothesis." *The Economic Journal* 128 (612): F404–F421.

**Vosters, Kelly, and Martin Nybom.** 2017. "Intergenerational Persistence in Latent Socioeconomic Status: Evidence from Sweden and the United States." *Journal of Labor Economics* 35 (3): 869–901.

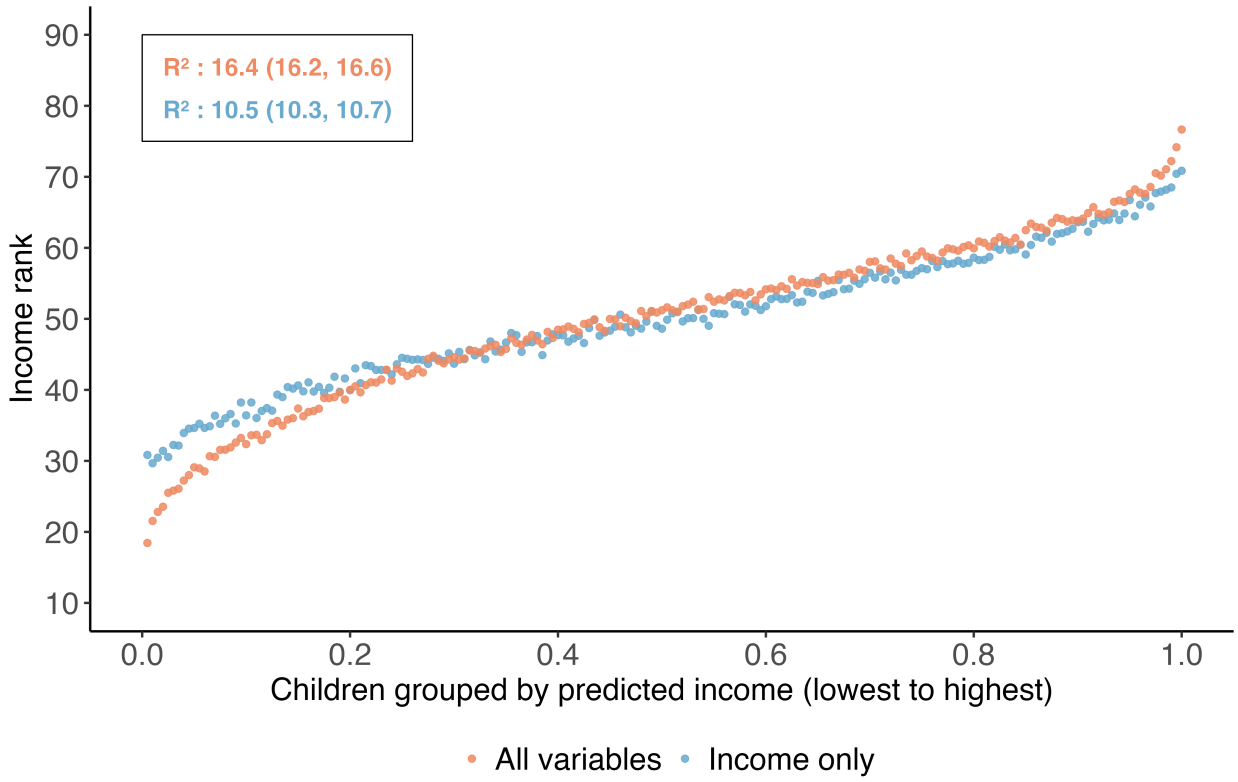
# Main Tables and Figures

Table 1: Explanatory Variables

| | |
|---|---|
| Income | Household income, personal income, personal earnings, most important sources of personal income (in 11 categories), and the primary household income share. |
| Wealth | The value of bank and savings balances, bonds and shares, real estate, entrepreneurial assets and liabilities, other assets, mortgage debt, study debt, and other debt. |
| Occupation | Average hourly wage and most important sector of employment (in 68 categories). |
| Education | Highest level of completed education. |
| Health | Average healthcare costs for 5 categories*: general practitioner, hospital, pharmaceutical, mental health care, and dental care. |
| Crime | Indicators of whether the parent has been suspected of a property, violent, or other type of crime. |
| Family structure | Parents' family size, age-at-first-birth, birth order, single-parent household, father or mother presence, parental death, child family size, and whether the father or the mother are identified. |
| Extended family outcomes | Average years of education, household income rank, wealth rank, total healthcare costs, and share of all siblings of the parent who have been suspected of a crime. |

Notes: this table describes the explanatory variables used in the main analysis. A detailed explanation of each of the variables and descriptive statistics can be found in Appendix B.
*: Healthcare costs are based on healthcare insurance reimbursements. Basic healthcare insurance is mandatory for all residents and covers a wide range of medical services (see also Appendix B).

Figure 1: Predicting Child Income with Detailed Parental Information



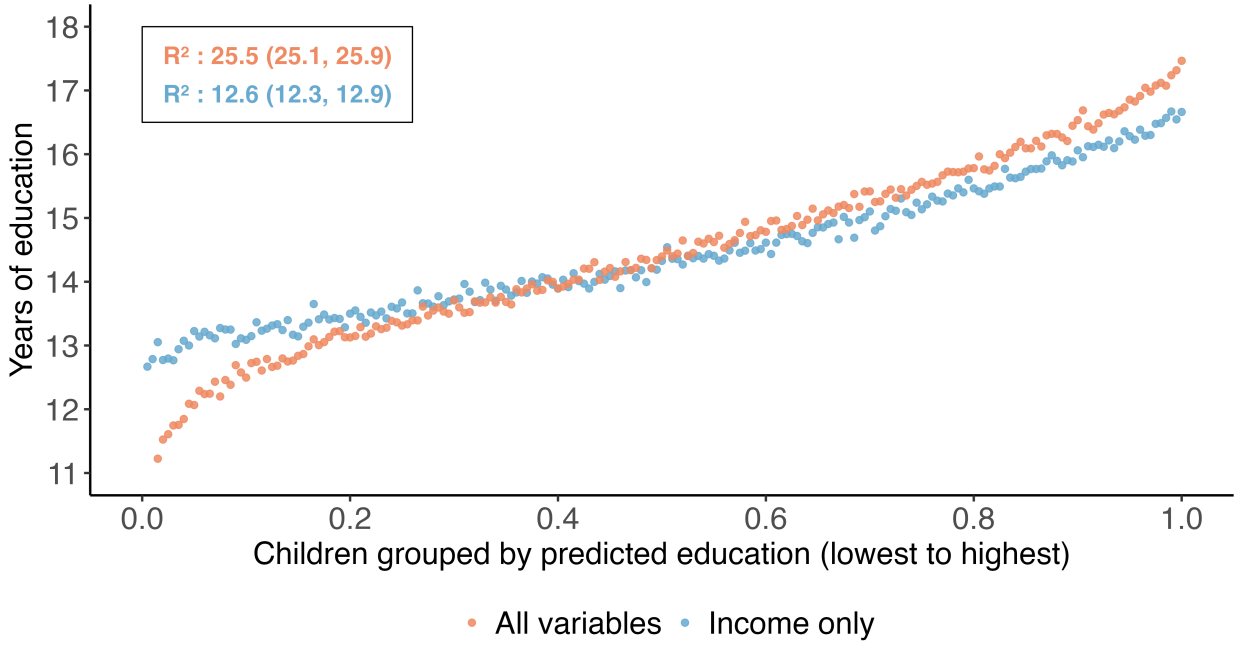

Notes: this figure presents binscatter plots of income ranks for 340,608 children in the test data, who are sorted into bins based on their predicted income rank according to two models. Both models are trained to predict children's income ranks using the same training sample of 1,362,430 children but include different explanatory variables. The orange graph is constructed as follows: (i) predict the income ranks of all children in the test data using the model with all explanatory variables, (ii) rank the predictions from low (0) to high (1) within a child's cohort, (iii) sort all children into 200 equal-sized bins based on their ranking, and (iv) calculate the average income ranks within each bin. The blue graphs are constructed similarly using the predictions from the model that uses parents' income only. Confidence intervals for the $R^2$ are bootstrapped from the test data using 599 draws.

Table 2: Family Background Characteristics across the Predicted Income Distribution

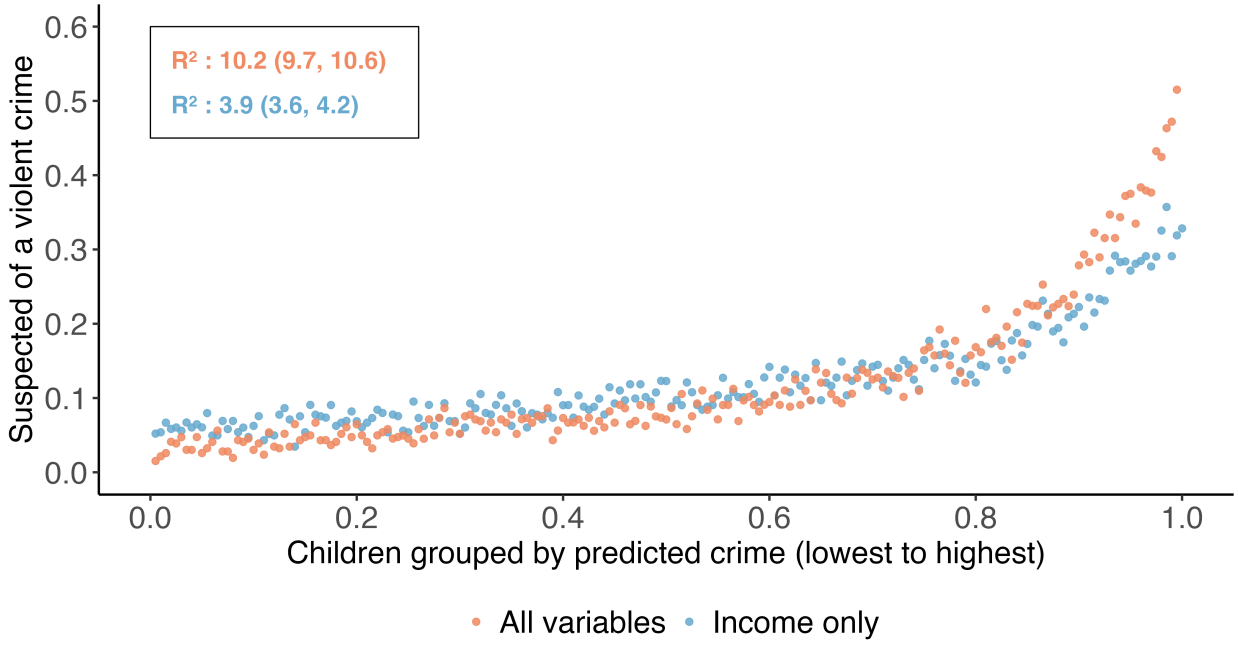
| | *Predicted Income Bins* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0-0.5 | 0.5-1 | 1-5 | 5-10 | 10-90 | 90-95 | 95-99 | 99-99.5 | 99.5-100 |
| Child income rank | 18.8 | 21.11 | 25.98 | 31.26 | 50.53 | 65.54 | 69.56 | 73.6 | 77.12 |
| | *Family background characteristics* | | | | | | | | |
| Parental income rank | 6.17 | 8.09 | 11.54 | 16.19 | 49.25 | 87.54 | 93.17 | 96.97 | 98.51 |
| Parental wealth rank | 11.82 | 12.66 | 14.21 | 17.18 | 50.86 | 74.29 | 80.43 | 86.53 | 90.04 |
| Max. education parents | 8.10 | 8.77 | 9.39 | 9.92 | 13.09 | 16.06 | 16.74 | 17.26 | 17.49 |
| Health costs parents | 5221 | 5064 | 4217 | 3862 | 2600 | 1873 | 1803 | 1542 | 1618 |
| Crime father | 0.59 | 0.47 | 0.32 | 0.18 | 0.05 | 0.02 | 0.02 | 0.03 | 0.02 |
| Extended family income | 17.02 | 20.4 | 25.15 | 30.43 | 49.13 | 64.72 | 69.32 | 74.53 | 78.65 |
| Extended family wealth | 21.78 | 24.04 | 26.74 | 30.71 | 51.1 | 63.85 | 67.6 | 70.5 | 73.81 |
| Father presence | 0.33 | 0.34 | 0.45 | 0.63 | 0.88 | 0.97 | 0.98 | 0.98 | 0.98 |
| Age at first birth mother | 21.95 | 22.76 | 23.93 | 25.23 | 27.07 | 28.39 | 28.6 | 28.78 | 29.05 |
| N | 1,703 | 1,703 | 13,624 | 17,030 | 272,487 | 17,030 | 13,624 | 1,703 | 1,704 |

Notes: Each column shows descriptive statistics for a group of children in the test data from the same predicted income bin. All values are averages, with missing values excluded from the calculations. The predicted income bins are constructed by predicting the income ranks of all children in the test data using the model with all explanatory variables, ranking them from low to high, and sorting them into bins according to their position in the predicted income distribution. Health expenditures parents equals the average health expenditures of the father and mother between 2009 and 2011. Extended family income (wealth) is calculated as the average income (wealth) rank of the father's and mother's siblings. The other variables are discussed in Table 1.

Figure 2: Predicting Children's Education and Crime




(a) Education




(b) Crime

Notes: the figures above present binscatter plots of children's years of education and crime for two predictive models. The children are sorted in 200 bins from lowest (0) to highest (1) predicted education/crime. Panel (a) reports results for 180,829 children from the test sample. Panel (b) reports the results for 92,725 sons from the test sample. The orange and blue dots are constructed using the same steps as in Figure 1. Confidence intervals for the $R^2$ are bootstrapped from the test data using 599 samples and are reported in brackets.

Table 3: Predictions within Neighborhoods, Migrant Groups, and Families with Adoptees

| | Income rank ($y$) | | | | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Predicted income ($\hat{y}$) | 1.003 | 0.939 | 0.999 | 0.939 | 0.273 | 0.874 |
| | (0.004) | (0.005) | (0.005) | (0.012) | (0.029) | (0.047) |
| *Fixed Effects* | | | | | | |
| Neighborhood | | x | | | | |
| Migration background | | | x | x | | |
| N | 333,792 | 333,792 | 333,792 | 68,889 | 4,935 | 3,802 |
| Sample | All | All | All | $2^d/3^d$ generation migrants | Adoptees | Own-birth children in adoption families |

Notes: Each column shows results from a separate regression of a child's income rank on its predicted value, applied to specific subsets of the data and/or including fixed effects. The predicted incomes are based on the gradient-boosted decision trees reported in figure 1. The samples in columns 1 to 6 correspond to the following children from the test data: (1-3) children with an available neighborhood identifier (97%), (4) second- and third-generation migrants, (5) international adoptees, and (6) own-birth children from families with at least one adopted child. Standard errors, shown in parentheses, are clustered at the neighborhood or migration background level in columns 2 to 4.

# Appendix A: supplementary results

Table A1: Descriptive statistics for the income analysis sample

| | Mean | SD | Mean | SD | % missing |
|---|---|---|---|---|---|
| **Characteristics children** | | | | | |
| Year of birth | 1984.6 | 2.9 | | | 0 |
| Male | 0.51 | 0.50 | | | 0 |
| Family size | 2.70 | 1.32 | | | 0 |
| Household income | 102,156 | 65,404 | | | 0 |
| Second generation migrant | 0.150 | 0.357 | | | 0 |
| Third generation migrant | 0.057 | 0.232 | | | 0 |
| | | | | | |
| **Family characteristics: measured at the household level** | | | | | |
| Household income rank | 0.499 | 0.289 | | | 0.009 |
| Primary income share | 0.794 | 0.268 | | | 0.011 |
| Highest education | 12.937 | 3.637 | | | 0.358 |
| Total wealth rank | 0.501 | 0.286 | | | 0.008 |
| Bank and savings balances | 52,249 | 180,944 | | | 0.008 |
| Bonds and shares | 36,704 | 347,226 | | | 0.008 |
| Substantial interest | 65,601 | 1,235,768 | | | 0.008 |
| House value | 309,747 | 379,964 | | | 0.008 |
| Entrepreneurial assets | 15,028 | 132,290 | | | 0.008 |
| Other real estate | 30,253 | 277,509 | | | 0.008 |
| Other assets | 6,090 | 111,069 | | | 0.008 |
| Total debt | 159,239 | 374,080 | | | 0.007 |
| Mortgage debt | 134,709 | 190,726 | | | 0.008 |
| | | | | | |
| **Relationship status of household head(s) of child at age 15:** | | | | | |
| Registered partners | 0.824 | 0.381 | | | 0.023 |
| Non-registered partners | 0.037 | 0.190 | | | 0.023 |
| Single parent | 0.126 | 0.332 | | | 0.023 |
| Other | 0.012 | 0.110 | | | 0.023 |
| | | | | | |
| **Other family characteristics** | **Father** | | **Mother** | | |
| Personal income | 68,129 | 51,443 | 29,157 | 21,734 | 0.108 |
| Personal earnings | 83,082 | 61,812 | 33,161 | 26,958 | 0.180 |
| | | | | | |
| *Most important source of income* | | | | | |
| Employment | 0.669 | 0.416 | 0.536 | 0.433 | 0.055 |
| Bonds or shares | 0.043 | 0.179 | 0.012 | 0.09 | 0.055 |
| Substantial interest | 0.005 | 0.051 | 0.03 | 0.123 | 0.055 |
| Entrepreneurship | 0.116 | 0.288 | 0.066 | 0.218 | 0.055 |
| Unemployment benefits | 0.025 | 0.091 | 0.017 | 0.062 | 0.055 |

| | | | | | |
|---|---|---|---|---|---|
| Welfare benefits | 0.022 | 0.132 | 0.046 | 0.187 | 0.055 |
| Disability insurance | 0.079 | 0.237 | 0.064 | 0.212 | 0.055 |
| Other security transfers | 0.004 | 0.049 | 0.007 | 0.062 | 0.055 |
| Pension | 0.023 | 0.109 | 0.037 | 0.147 | 0.055 |
| Other sources | 0.014 | 0.087 | 0.185 | 0.338 | 0.055 |
| | | | | | |
| *Type of housing* | | | | | |
| Own house | 0.745 | 0.409 | 0.700 | 0.428 | 0.066 |
| Rental | 0.053 | 0.190 | 0.104 | 0.259 | 0.066 |
| Subsidized rental | 0.200 | 0.356 | 0.195 | 0.338 | 0.066 |
| | | | | | |
| Years of education | 12.79 | 3.83 | 11.93 | 3.6 7 | 0.530 |
| Average hourly wage | 32.0 | 26.9 | 20.7 | 18.1 | 0.315 |
| Most important sector of employment | - | - | - | - | 0.315 |
| Suspected of any crime | 0.067 | 0.25 | 0.023 | 0.15 | 0.014 |
| Suspected of property crime | 0.014 | 0.119 | 0.008 | 0.09 | 0.014 |
| Suspected of violent crime | 0.025 | 0.156 | 0.006 | 0.079 | 0.014 |
| Suspected of other crime | 0.042 | 0.2 | 0.012 | 0.11 | 0.014 |
| Total health costs | 2700 | 7153 | 2626 | 8212 | 0.063 |
| General practitioner costs | 174 | 143 | 197 | 155 | 0.063 |
| Mental health care costs | 234 | 3541 | 321 | 3948 | 0.063 |
| Hospital care costs | 1830 | 6723 | 1692 | 5013 | 0.063 |
| Pharmaceutical care costs | 527 | 2230 | 542 | 2084 | 0.063 |
| Dental care costs | 46 | 303 | 44 | 299 | 0.063 |
| | | | | | |
| Year of birth | 1953.6 | 5.6 | 1956.4 | 5.0 | 0.009 |
| Age at first birth | 29.3 | 5.5 | 27.0 | 4.4 | 0 |
| Family size | 4.1 | 2.4 | 4.0 | 2.3 | 0.218 |
| Birth order | 2.5 | 1.8 | 2.5 | 1.8 | 0.218 |
| Father/mother identified | 0.025 | 0.157 | 0.002 | 0.049 | 0 |
| Father/mother dead | 0.008 | 0.086 | 0.004 | 0.065 | 0.019 |
| Father/mother present in household | 0.857 | 0.35 | 0.962 | 0.191 | 0.037 |
| | | | | | |
| *Extended family outcomes* | | | | | |
| Average income rank | 0.496 | 0.222 | 0.495 | 0.224 | 0.246 |
| Average education | 12.61 | 3.155 | 12.732 | 3.103 | 0.420 |
| Average wealth rank | 0.514 | 0.226 | 0.511 | 0.227 | 0.239 |
| Average health expenditures | 2717 | 5537 | 2564 | 5370 | 0.231 |
| % siblings suspected of any crime | 0.043 | 0.142 | 0.048 | 0.153 | 0.231 |

*Note:* This table presents descriptive statistics of the income sample. The sample comprises of all $n = 1,703,038$ children born between 1980 and 1989 with non-missing income (96.6%). A detailed explanation of the variables can be found below this table.

**Income.** The construction of children's and parents' household income ranks is discussed in the main text.

The share of primary income represents the fraction of household income derived from labor, entrepreneurship, or capital. It is constructed similarly to parental household income. Specifically, for each parent, I calculate the primary income share for each year up to age 60—the same years in which household income is measured. The lifetime primary income share is then defined as the average of these yearly shares. Finally, the household share of primary income is determined by averaging the lifetime primary income shares of both parents.

Personal income refers to an individual's income from labor, entrepreneurship, or transfers, measured at the personal rather than household level. As a result, it excludes partners' incomes but also household-level income streams, such as capital gains or rental allowances. Personal earnings equals personal income minus income transfers. Following the same approach as before, I exclude years with income or earnings observations lower than €1000, and proxy a parent's lifetime personal income and earnings by averaging all personal income and earnings observations up to age 60. Although the table above shows personal income and earnings in absolute values, in the analysis, I use ranks instead. The ranks are taken relative to all other parents in the sample.

In addition, I identify the primary sources of personal income, classified into 10 categories.[24] Drawing on all yearly observations used in constructing the lifetime personal income measure, I first compute the most important source of income in each of those years. I then compute the fraction of years in which each category served as the main source of income.

Similarly, for each of those years, I calculate the fraction of years that the father or the mother lived in a self-owned house, a rental property, or a government-subsidized rental.

**Wealth.** The wealth variables are constructed in a manner analogous to the parental household income variable, as both are measured at the household level. I observe the values for each type of asset or liability of each parent in 2006. Subsequently, for each child, I determine the mean of the father's and mother's values for each asset or liability type.

The assets and liabilities included in this analysis are defined as follows. Bank and savings balances represent the total deposits held by a household in (savings) bank accounts, including foreign accounts. House value captures the market value of a household-owned dwelling used as the primary residence, while other real estate encompasses the total value of any additional properties owned by the household. Bonds and shares measure the combined value of bond and equity holdings, excluding 'substantial interests' (holdings of at least 5 percent of a company's issued share capital), which are accounted for separately under the "substantial interests" variable. Entrepreneurial assets reflect the net balance of a household's business-related assets and liabilities, and other assets include any remaining assets not covered by the aforementioned categories. Mortgage debt refers to debts associated with the household's owner-occupied home, whereas other debt encompasses all other types of liabilities.

[24] One category is income from substantial interest. A substantial interest refers to a shareholder owning at least 5% of a company's shares. This threshold is used for tax and regulatory purposes to identify large or influential shareholders. Income and wealth from such shares are measured separately.

**Education.** Parents' years of education are based on the conversion table in Appendix E1. Table A1 indicates that parental education information is absent for about 50 percent of the sample. This gap exists because Statistics Netherlands initiated systematic education data collection only in the late 1980s. Prior educational records are mainly sourced from large-scale surveys frequently administered by Statistics Netherlands and are also obtained indirectly from other government bodies, including the unemployment agency.

**Occupation.** I use monthly data on all employment contracts in the Netherlands from 2006 to 2009, collected by the tax authorities through third-party reporting. For each individual, I aggregate the total hours worked at each firm during this period. I then identify the firm where the individual has accumulated the most hours and assign the individual's employment sector based on that firm's classification. Sector categorizations are determined by the authorities in accordance with collective labor agreements. There are 68 sector categories in total, which include categories such as 'education and sciences', 'government defense', 'chemical industry', 'financial services', 'restaurants and bars', 'retail', etc. The average hourly wage is calculated by dividing the individual's total gross salary over the period by the total number of hours worked.

**Health.** The health care expenditures are based on annual healthcare costs for care covered by the basic insurance. The basic insurance is legally mandated under the Healthcare Insurance Act for nearly all residents of the Netherlands. The costs refer to expenses for all types of care that are reimbursed by health insurers, and may include amounts ultimately paid by the insured themselves due to the deductible, but exclude copayments. If the insured received a bill and did not submit it to the insurer—e.g., because the deductible had not been reached—these costs are not included in the figures. The health care expenditures variables above are based on the subcategories of healthcare spending defined by Statistics Netherlands. For each of the subcategories, the annual costs are averaged over the period 2009 to 2011.

**Crime.** As explained in section 2, the crime data contains all offenses reported to the police between 2005 and 2022. The data contain the reporting date, the offense type, and the individual identifier of the suspected offender(s) whenever there is a known suspect. I use these data to construct indicators of whether the father or the mother has been suspected of different types of crimes between 2005 and 2010.

**Family structure.** I record the family size and birth order of both the father and the mother by linking them to their siblings, which requires accessing the grandparents' identifiers. Consequently, these variables, along with any extended family outcomes, are missing for children whose grandparents cannot be identified. Additionally, I determine whether the father or mother was registered in the same household as the child at age 15 and classify the child's household type at that age into one of three categories: a couple with a registered partnership, a couple without a registered partnership, or a single-parent household. Furthermore, I calculate the parents' age at the birth of their first child and indicate whether either the father or the mother is not identified, as not all children have both parents identified.

**Extended family outcomes.** For each parent separately, I determine the mean years of education, household income rank, wealth rank, and annual health expenditures across all their siblings. Additionally, I calculate the fraction of these siblings who have been suspected of committing a crime.

Table A2: Predicting child income using smaller samples

| Share of core sample | Test data sample size | $R^2$ | 0.025% lower bound | 97.5% upper bound |
|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) |
| 0.01 | 3,406 | 0.137 | 0.116 | 0.161 |
| 0.02 | 6,812 | 0.147 | 0.130 | 0.165 |
| 0.05 | 17,031 | 0.149 | 0.140 | 0.158 |
| 0.1 | 34,061 | 0.157 | 0.151 | 0.165 |
| 0.2 | 68,122 | 0.158 | 0.152 | 0.163 |
| 0.4 | 136,243 | 0.162 | 0.159 | 0.166 |
| 0.6 | 204,365 | 0.162 | 0.159 | 0.165 |
| 0.8 | 272,486 | 0.162 | 0.159 | 0.164 |

Notes: This table presents estimates of explanatory power for gradient-boosted decision trees that include all explanatory variables (as in Figure 1), using smaller samples. Column 1 reports the share of the core sample that is used for the analysis. Column 2 reports the sample size of the test-data. Columns 3, 4, and 5 report the $R^2$ and 95% confidence interval lower and upper bounds, respectively. Each model is trained on a randomly selected 80% of the respective sample, and evaluated on the remaining 20%. Confidence intervals for the $R^2$ are bootstrapped from the test-data using 599 draws.

Table A3: Predicting child income: varying years and ages of income measurement

| | $R^2$ | 0.025% lower bound | 97.5% upper bound |
|---|---|---|---|
| Years of income | A. Varying years of income measurement | | |
| 1 | 0.137 | 0.133 | 0.141 |
| 2 | 0.144 | 0.141 | 0.148 |
| 3 | 0.150 | 0.146 | 0.154 |
| 4 | 0.152 | 0.148 | 0.156 |
| 5 | 0.156 | 0.152 | 0.161 |
| 6 | 0.157 | 0.153 | 0.161 |
| 7 | 0.160 | 0.156 | 0.164 |
| 8 | 0.160 | 0.156 | 0.164 |
| 9 | 0.163 | 0.159 | 0.167 |
| All | 0.168 | 0.164 | 0.172 |
| Age child | B. Varying ages of income measurement | | |
| 30-33 | 0.127 | 0.123 | 0.131 |
| 34-37 | 0.154 | 0.149 | 0.158 |
| 38-41 | 0.153 | 0.149 | 0.157 |
| All > age 32 | 0.166 | 0.162 | 0.170 |

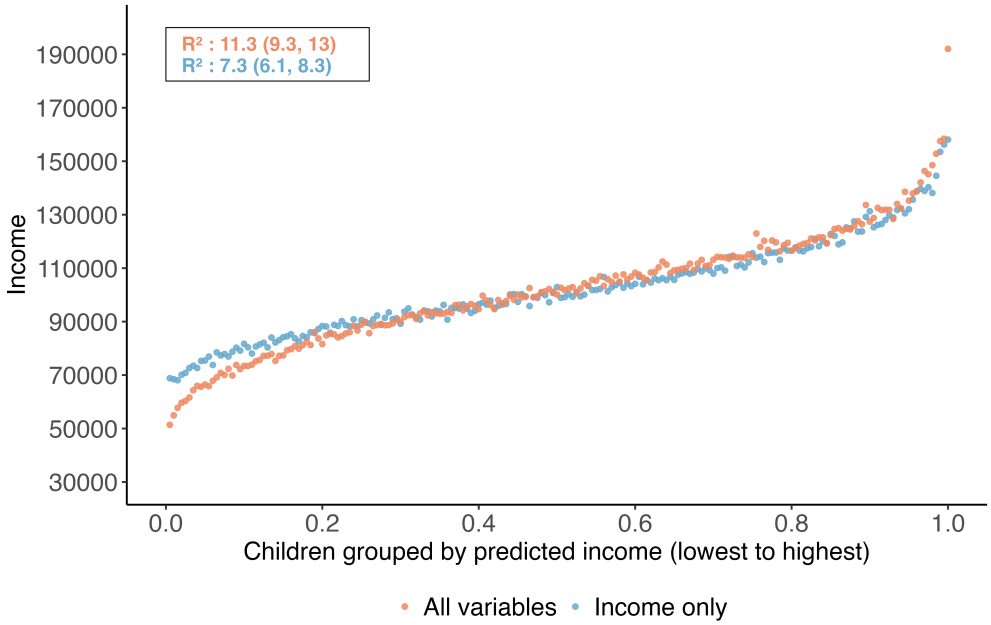
Notes: Each row presents the $R^2$ and corresponding 95% lower and upper bound for gradient-boosted decision trees that include all explanatory variables to predict child income (as in Section 4). The analysis sample consists of all 330,018 children born in 1980 and 1981 for whom I observe all incomes between ages 30 and 41. Each model is trained on the same randomly selected 80% of this sample, and evaluated on the remaining 20%. Panel A varies the number of years of income data used to construct the child income rank. The last row in panel A uses all income observations, as in the main results. Panel B rows 1 to 3 use four years of income data, but vary the ages at which income is measured. Row 4 uses all income data above age 32. Confidence intervals for the $R^2$ are bootstrapped from the test-data using 599 draws.

Table A4: The effect of family background on income: regression results with adoptees

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| (5) | | | | |
| Predicted income | 0.273 | 0.274 | 0.277 | 0.271 |
| | (0.029) | (0.029) | (0.030) | (0.024) |
| Controls | | x | x | x |
| Country of Origin FE | | | x | x |
| Year of Adoption FE | | | | x |
| N | 4,935 | 4,935 | 4,935 | 4,935 |

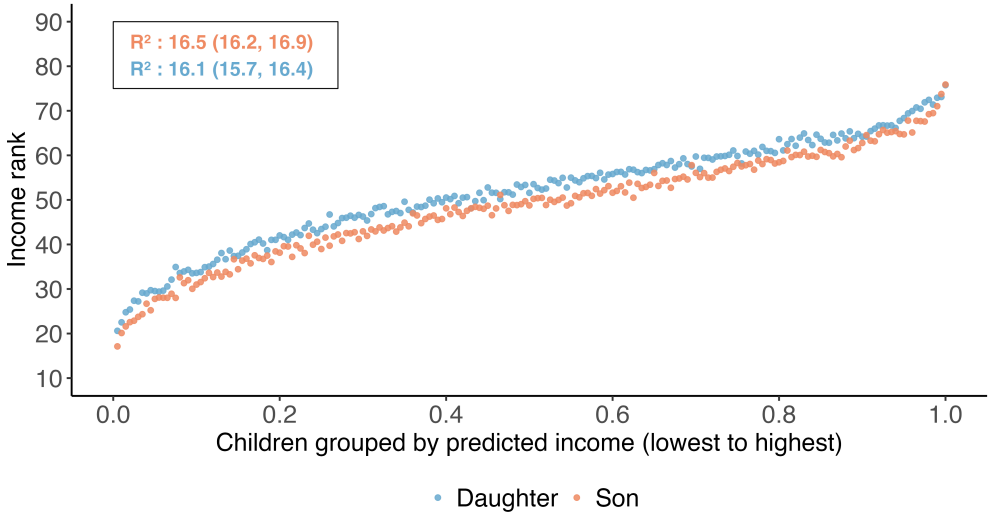
Notes: Each column shows results from a separate regression of an adopted child's income rank on predicted income. Predicted incomes are based on the calibrated machine learning model (as in Section 4). Controls are a gender dummy and age-at-migration. The predicted values for income are based on gradient-boosted decision trees reported in Figure 1. The fixed effects are fully interacted. Standard errors are in parentheses. (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

Figure A1: Predicting children's income level



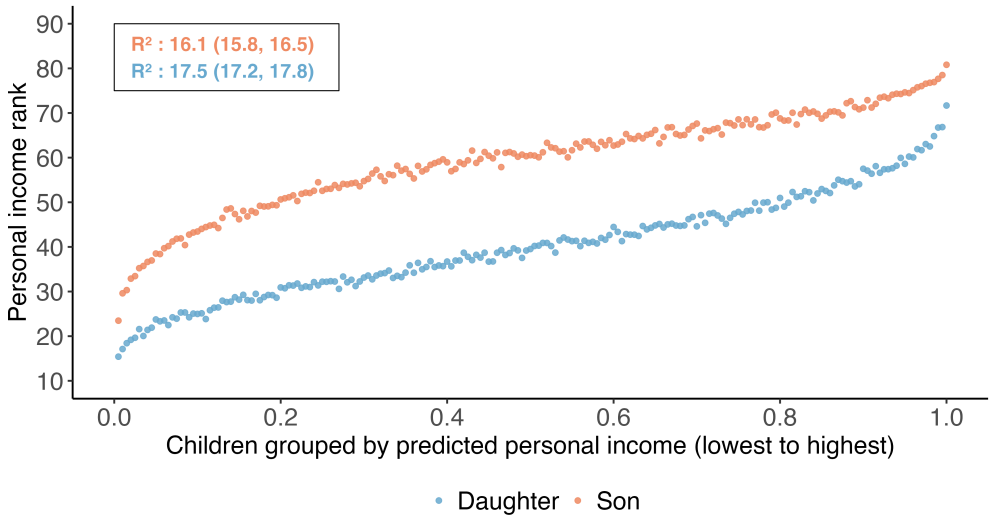

Notes: this figure presents binscatter plots of children's household income, who are sorted into bins based on their predicted income rank. The graphs are constructed using the same sample and steps as in Figure 1, applied to children's income levels instead of ranks. Confidence intervals for the $R^2$ are bootstrapped from the test data using 599 samples and are reported in brackets

Figure A2: Predicting children's household income rank by gender




Notes: this figure presents binscatter plots of sons' and daughters' household income ranks for 173,652 sons and 166,957 daughters in the test data, who are sorted into bins based on their predicted income rank. Predictions are generated using the same predictive model and explanatory variables as in Section 4, now applied separately to each gender. The construction of the graphs follows the same steps as in Figure 1, now separately for each gender. Confidence intervals for the $R^2$ are bootstrapped from the test data using 599 samples and are reported in brackets

Figure A3: Predicting children's personal income by gender




Notes: this figure presents binscatter plots of sons' and daughters' personal income ranks for 172,976 sons and 164,990 daughters in the test data, who are sorted into bins based on their predicted income rank. The graphs are constructed using the same steps as in Figure 1, applied to children's personal income ranks instead of household income ranks. Confidence intervals for the $R^2$ are bootstrapped from the test data using 599 samples and are reported in brackets

Table A5: Descriptive statistics for international adoptees and their parents

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted income rank ($\hat{y}$) | 38.10 | 45.67 | 49.21 | 51.93 | 54.35 | 56.59 | 58.75 | 60.94 | 63.83 | 69.22 |
| Income rank ($y$) | 35.94 | 37.17 | 40.24 | 37.79 | 42.21 | 41.33 | 40.79 | 43.87 | 42.77 | 43.06 |
| | *Characteristics Adoptive Parents* | | | | | | | | | |
| Parental income rank | 19.81 | 29.6 | 37.71 | 43.89 | 52.64 | 59.65 | 68.4 | 76.61 | 85.22 | 93.13 |
| Parental wealth rank | 31.40 | 45.95 | 53.26 | 58.69 | 63.34 | 67.67 | 68.8 | 71.10 | 71.49 | 81.57 |
| Highest education parents | 11.19 | 11.85 | 12.80 | 13.67 | 14.21 | 14.41 | 15.20 | 15.30 | 16.03 | 16.38 |
| Father suspected of crime | 0.10 | 0.05 | 0.04 | 0.05 | 0.03 | 0.03 | 0.01 | 0.02 | 0.02 | 0.02 |
| Health expenditures parents | 4,610 | 3,883 | 3,276 | 2,942 | 3,488 | 2,401 | 2,847 | 2,622 | 2,589 | 2,032 |
| Extended family income rank | 38.12 | 43.49 | 49.29 | 49.84 | 52.52 | 56.13 | 57.14 | 61.42 | 61.93 | 70.34 |
| N | 493 | 494 | 493 | 494 | 493 | 494 | 493 | 494 | 493 | 494 |

Notes: Each column shows descriptive statistics for a group of international adoptees from the same predicted income bin. All cells are averages. The predicted income bins are constructed by predicting the income ranks of all adoptees using the model with all explanatory variables, ranking them from low to high, and sorting them into ten equally sized bins according to their position in the predicted income distribution.

# Appendix B: intergenerational mobility estimates

This appendix briefly discusses additional intergenerational mobility estimates to evaluate the sensitivity of the rank-rank correlation of 0.32 to various specification choices. Although it would be ideal to perform robustness checks using the full analysis sample, the specific data requirements for each check necessitate the use of different samples. Stability of the estimates within these samples strengthens confidence that the estimates would also remain stable under different specifications in the broader analysis sample.

First, Figure A4 presents a binscatter plot of children's income ranks against parental income ranks for the full sample. This plot provides a non-parametric description of how children's incomes vary by their parents' incomes.

Table B1 reports the rank-rank correlation as well as the Intergenerational Income Elasticity (IGE) using logs of household income instead of ranks for the full analysis sample in columns 1 and 2. These are, coincidentally, equal up to the second digit. Columns 3 and 4 report results for sons and daughters separately and rely on children's personal income ranks instead of household income ranks. These estimates are very similar and close to the rank-rank correlation using the pooled sample and household income.

Table B2 reports mobility estimates using varying years of income information for both parents and children. I focus on all children born in 1985 because for this group I have the highest income data availability of both parents and children, allowing me to analyze the sensitivity. The estimates attenuate somewhat with fewer years of income, but the change in the rank-rank correlation is limited after 5 years of income are used. In the core sample, I have at least 5 years of income observation for almost all children and parents.

Table B3 reports mobility estimates using incomes of children measured at varying ages. I focus on all children born in 1980 or 1981 for whom all incomes are observed between ages 30 to 41. I average income over 4 years for each of the specifications. The estimates show that measuring income early attenuates the estimates, but they stabilize after age 34. Overall, the differences are relatively small.

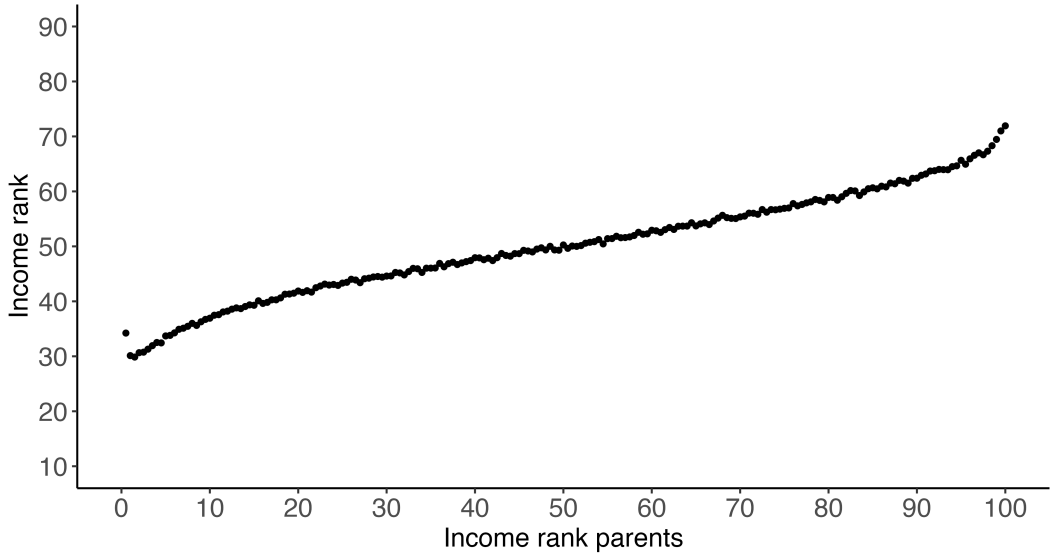
Finally, Table B4 reports mobility estimates using incomes of parents measured in different periods. I focus on all children for whom parental income is observed between 2003 and 2013. I average income over 5 years for each of the specifications. The estimates are very similar, regardless of when parental income is measured.

Table B1: Intergenerational mobility estimates

| | Rank rank correlation | IGE | Personal income rank (daughters) | Personal income rank (sons) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Coefficient | 0.323 | 0.324 | 0.288 | 0.290 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| N | 1,702,355 | 1,702,355 | 864,064 | 825,170 |
| $R^2$ | 0.105 | 0.091 | 0.093 | 0.095 |

Notes: column (1) shows results from a regression of a child's household income rank on the parents' household income rank. Column (2) shows results from a regression of the log of child household income on the log of parental household income. Columns (3) and (4) show results from a regression of sons' or daughters' personal income rank on parents' household income rank. The sample includes the core analysis sample (Table A1) excluding observations with missing parental income (0.9 percent). Standard errors are in parentheses.

Figure A4: Mean Child Income Rank vs. Parent Income Rank




Notes: this figure presents a nonparametric scatter plot of mean income ranks versus parental income rank. The sample consists of $N = 1,702,355$ children. The $X$-axis reports the parent income rank sorted into 200 equal-sized bins. The $Y$-axis reports the mean income rank within each bin.

Table B2: Intergenerational mobility estimates: varying years of income

| Years of income | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Coefficient | 0.276 | 0.288 | 0.301 | 0.306 | 0.312 | 0.316 | 0.320 | 0.322 | 0.324 |
| | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) | (0.002) |
| N | 169,504 | 169,504 | 169,504 | 169,504 | 169,504 | 169,504 | 169,504 | 169,504 | 169,504 |
| $R^2$ | 0.076 | 0.083 | 0.091 | 0.094 | 0.098 | 0.100 | 0.103 | 0.104 | 0.105 |

Notes: each column presents results from a regression of a child's household income rank on the parents' household income rank. The number of years of income data used to construct the income rank varies across columns, as indicated in the first row. The income observations used are always those closest to age 35. Standard errors are reported in parentheses. The sample consists of all children from the core sample born in 1985.

Table B3: Intergenerational mobility estimates: measuring child income at different ages

| | (1) | (2) | (3) |
|---|---|---|---|
| Coefficient | 0.274 | 0.304 | 0.308 |
| | (0.002) | (0.002) | (0.002) |
| Age child | 30-33 | 34-37 | 38-41 |
| N | 326,388 | 326,388 | 326,388 |
| $R^2$ | 0.076 | 0.093 | 0.095 |

Notes: Each column presents results from a regression of a child's household income rank on the parents' household income rank. Parent household income is measured as in the main results of this paper. Child household income ranks are always based on 4 years of income, but the ages at which child incomes are measured vary across columns. The sample consists of all children born in 1980 or 1981 for whom all incomes between ages 30 and 41 are available. Standard errors are reported in parentheses.

Table B4: Intergenerational mobility estimates: measuring parent income at different ages

| | (1) | (2) | (3) |
|---|---|---|---|
| Coefficient | 0.290 | 0.294 | 0.292 |
| | (0.001) | (0.001) | (0.001) |
| Years of income measurement parents | 2003-2007 | 2006-2010 | 2009-2013 |
| N | 1,267,606 | 1,267,606 | 1,267,606 |

Notes: Each column presents results from a regression of a child's household income rank on the parents' household income rank. Child income ranks are measured as in the main analysis in this paper. Parent household income ranks are always based on 5 years of income, but the periods at which incomes are measured vary across columns. The sample consists of all children in the core sample for whom parental income is observed between 2003 and 2013. Standard errors are reported in parentheses.

# Appendix C: Measuring variable importance

Interpreting gradient-boosted decision trees is notoriously difficult due to their complexity. However, gaining insight into which variables add most explanatory power is highly valuable. Recent advances in machine learning now allow us to compute the contribution of each variable to specific predictions using Shapley values. Below, I provide a brief explanation of the intuition behind this approach, followed by a graph displaying the Shapley values for the 30 most predictive variables in the analysis.

Shapley values originate from cooperative game theory (Shapley (1953)). In this framework, a coalition of agents $j \in S$ produces an output $\nu(S)$. The Shapley value for agent $i \in S$ represents its average marginal contribution to the output $\nu(s)$ across all possible coalitions $s \subseteq S \backslash i$. This concept directly applies to prediction models, where the output $f(x_1, ..., x_k)$ is generated from a set of variables $x_j \in X^j$, with $j \in \{1, ..., k\}$. In this context, Shapley values represent the average marginal contribution of each variable to a prediction, calculated by averaging over all possible subsets of included variables.[25]

Lundberg and Lee (2017) show that Shapley values are the only measures of variable importance that preserve important properties from cooperative game theory.[26] While exact Shapley values are computationally infeasible for most models due to the need to sum over all feature subsets (an NP-hard problem), recent algorithms can compute exact Shapley values for tree-based models in short time periods (Lundberg et al. (2020)).

Using this algorithm, I compute Shapley values for the gradient-boosted decision tree model used in the main results (1), applied to a random sample of 10,000 children from the test dataset. This process generates Shapley values for each variable and each child. Of all explanatory variables, only the indicators for whether the father or mother is identified in the data provide no contribution to the predictions. Figure C1 presents a boxplot of Shapley values for the 30 variables with the highest average absolute Shapley values, ranked from highest to lowest.

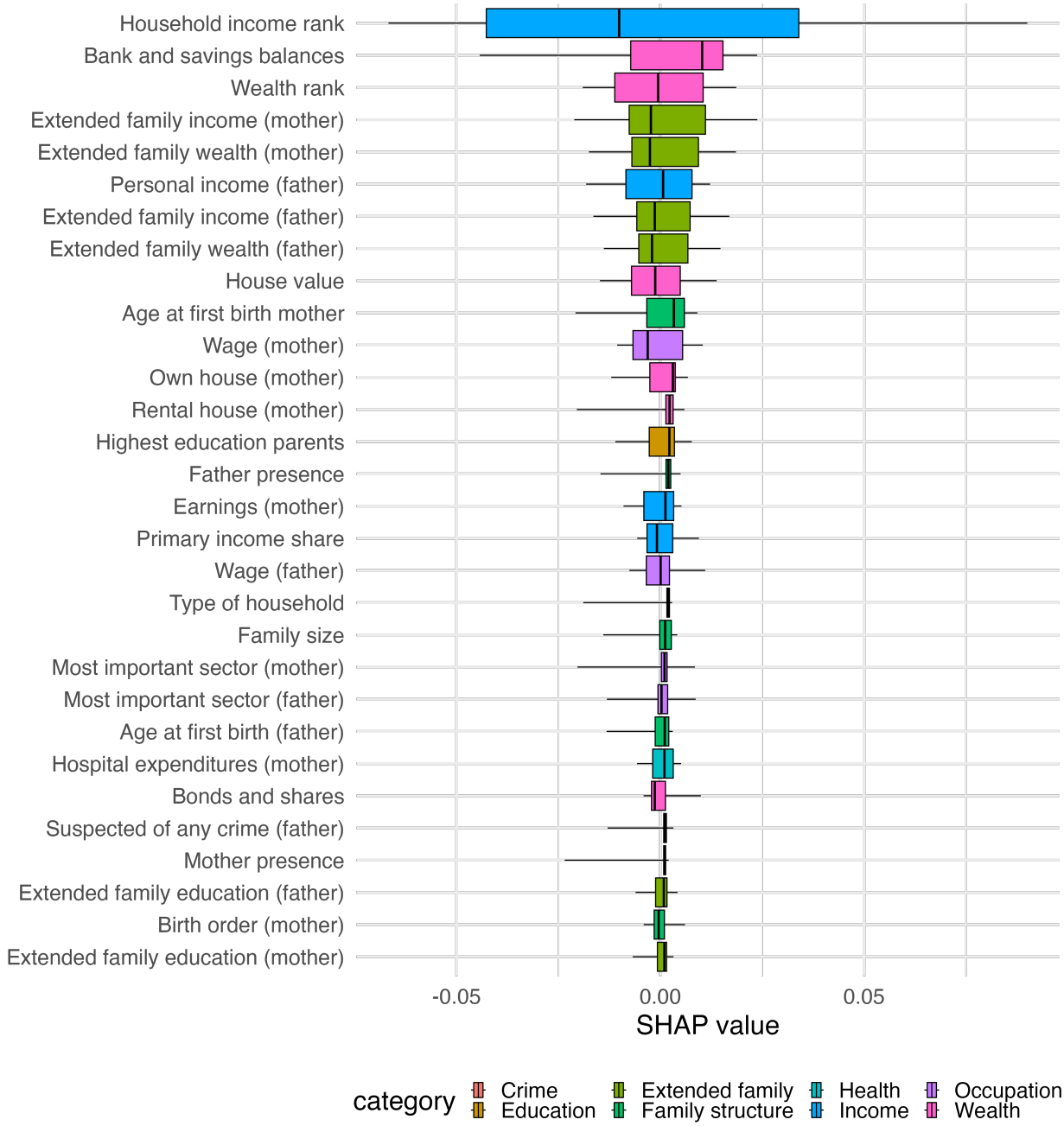
To illustrate, consider the boxplot for the household income rank: the 2.5th percentile is -0.07, indicating that for 2.5 percent of the children, parental income reduces the prediction by at least 0.07 ranks compared to the average prediction of 0.50. The 75th percentile is 0.03, meaning that for 25 percent of the children, parental income increases the prediction by more than 0.03 ranks relative to the average.

Figure C1 shows that the nine variables with the highest average absolute Shapley values are all related to parental or extended family income and wealth. The spread of the Shapley values for these variables is relatively large, which means that they provide sizeable contributions to the predictions for many children. Nonetheless, other variables also have meaningful impacts. For example, although mother's age-at-first-birth or mother presence have smaller average contributions, these variables exert a substantial negative effect on a small subset of children.

[25] This contrasts with commonly reported $R^2$-based Shapley values, which quantify variables' marginal contributions to the overall model fit ($R^2$), rather than to specific individual predictions.

[26] These properties are local accuracy and consistency. Local accuracy (additivity) ensures that for a given input $x$, the sum of the Shapley values equals the model's output $f(x)$. Consistency (monotonicity) guarantees that if a variable's contribution increases or stays the same, its Shapley value will not decrease, regardless of the other inputs.

Figure C1: Measuring variable importance using Shapley values




Notes: this figure presents boxplots of Shapley values for 30 explanatory variables. Shapley values are computed using the algorithm of Lundberg et al. (2020) for each variable and each child using a randomly drawn sample of 10,000 children from the test dataset. The variables shown are those with the 30 highest mean absolute Shapley values across these observations. Each row displays a boxplot representing the distribution of Shapley values for a given variable. The whiskers indicate the 2.5th and 97.5th percentiles, the box edges correspond to the 25th and 75th percentiles, and the center bar represents the mean. Explanatory variables are color-coded by category.

# Appendix D: a comparison of methods

This section presents a simple framework linking the approach in this paper to intergenerational mobility estimates, sibling correlations, and inequality of opportunity estimates.

Let $Y_{sf}$ be the income of a child $s$ in a family $f$ and let $Y_f$ be parental income. Let $\mathbf{X}_f = (Y_f, X_{f1}, ..., X_{fk}) \subset \mathcal{X}$ be the observable features that siblings share and $\mathbf{Z}_f = (Z_{f1}, ..., Z_{fl}) \in \mathcal{Z}$ be the unobservables features that siblings share.[27] Consider the following two conditional expectations function decompositions of $Y_{sf}$:[28]

1. *Sibling model*:

$$Y_{sf} = E[Y_{sf}|\mathbf{X}_f, \mathbf{Z}_f] + e_{sf} = f(\mathbf{X}_f, \mathbf{Z}_f) + e_{sf}, \tag{1}$$

2. *Observables model*:

$$Y_{sf} = E[Y_{sf}|\mathbf{X}_f] + \nu_{sf} = g(\mathbf{X}_f) + \nu_{sf}, \tag{2}$$

where $E[e_{sf}] = E[e_{sf}h(\mathbf{X}_f, \mathbf{Z}_f)] = 0$ for any $h : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ and $E[\nu_{sf}] = E[\nu_{sf}m(\mathbf{X}_f)] = 0$ for all $m : \mathcal{X} \to \mathbb{R}$.

Both models decompose income variation into mean differences between groups and residual variation within groups. In the sibling model, the groups consist of siblings, who by construction share both observable and unobservable features. In the observables model, the groups include all children sharing the same observable family characteristics.

The primary objective of this paper is to measure the importance of observable family background characteristics for children's income. I quantify this by the share of income variation attributable to differences in $g(\mathbf{X}_f)$ — the conditional mean for individuals with observable family background $\mathbf{X}_f$ - as opposed to residual variation in income $\nu_{sf}$. This corresponds to the non-parametric $R^2$ of the observables model:

$$R^2_{y|g} = \frac{V(g(\mathbf{X}_f))}{V(Y_{sf})}. \tag{3}$$

I commonly refer to this metric as the 'explanatory power'.

Fudenberg et al. (2022) show that a predictive model can explain a small amount of the variation in outcomes, and yet capture most of the *predictable* variation given the set of variables. There is an interesting analogy for the current setting. That is, the sibling-shared environment can have little explanatory power for income, which means that $f$ and $g$ will have little explanatory power. Still, however, $g$ can be a good approximation of $f$. In that

[27] I focus exclusively on variables that siblings share. As a result, parental factors differing between siblings, such as life-cycle variations in earnings or birth order effects, are excluded from the analysis. Restricting the model to variables that siblings share allows me to easily compare the results to sibling correlations.

[28] See, for example, Angrist and Pischke (2009) theorem 3.1.1 for a proof. The decompositions provide statistical associations and do not represent causal relationships.

case, the fraction of the variance in $f(\mathbf{X}_f, \mathbf{Z}_f)$ that is explained by $g(\mathbf{X}_f)$,

$$R^2_{f|g} = \frac{V(g(\mathbf{X}_f))}{V(f(\mathbf{X}_f, \mathbf{Z}_f))} = \frac{R^2_{y|g}}{R^2_{y|f}}, \tag{4}$$

is high. In the expression above, $R^2_{y|f} = V(f(\mathbf{X}_f, \mathbf{Z}_f))/V(Y_{sf})$ equals the sibling correlation. Even though $f$ relies on unobservables, $R^2_{y|f}$ is identified because its value coincides with the correlation between two randomly drawn siblings. A value of $R^2_{f|g}$ close to zero means that siblings' similarities arise from factors uncorrelated with the observables. On the other hand, if $R^2_{f|g}$ is close to one, then the observables are nearly as predictive as the model that includes unobservables $\mathbf{Z}_f$.[29]

An intergenerational mobility regression of $Y_{sf}$ on $Y_f$ represents a specific case of the broader observables model (2). It uses a subset of the observables - parental income only - and imposes a linearity restriction. Consequently, whereas the sibling correlation bounds the explanatory power of $g(\mathbf{X}_f)$ from above, the explanatory power of an intergenerational mobility regression is weakly lower than that of the full observables model. There is a one-to-one relationship between the slope of this regression, $\beta$, and its explanatory power: $R^2 = \beta^2 V(Y_f)/V(Y_{sf})$. As a result, intergenerational mobility coefficients are easily comparable to explanatory power estimates from sibling correlations or predictive models.

Finally, a closely related approach from the inequality of opportunity literature makes similar decompositions as in Equation 3, but typically uses other inequality measures than the variance. This is called the ex-ante approach to quantifying inequality of opportunity.[30] This literature uses multiple observable factors as explanatory variables, referred to as 'circumstances', which are beyond an individual's control. The findings in this paper are specific to inequality of opportunity arising from family circumstances, a subset of all possible circumstances.

[29] Standard decompositions of sibling correlations rely on strong linearity and homogeneity assumptions (Solon (1999)). An exception is Bingley and Cappellari (2019), who show that allowing for unobserved heterogeneity in transmission across families greatly increases the importance of parental influences. Instead of modeling *unobserved* heterogeneity, the decomposition above shows how flexible predictive models with many *observable* variables can be related to sibling correlations.

[30] A detailed explanation of this and related approaches can be found in Roemer and Trannoy (2016) and Ramos and Van de Gaer (2016). Brunori et al. (2024) also discuss how intergenerational mobility coefficients and inequality of opportunity estimates are related.

# Appendix E: a conversion table for years-of-education

For the educational outcome, I convert an individual's highest level of completed education into a years-of-education variable. Figure E1 provides a simplified overview of the levels of education and their corresponding years of schooling. The abbreviations are explained in Table E1. Generally, I convert the level of education into the number of years it takes to finish this type of education without delays. For example, an individual who has a university (WO) bachelor is assigned 17 years of education (8 years of primary school, 6 years of secondary education, and 3 years of university education). However, as indicated in Figure E1 by the downward arrow, more years of education does not necessarily imply a higher level. For example, it takes 16 years to obtain a vocational education (MBO) degree and 13 years to obtain a higher vocational secondary education (HAVO) degree, but both grant access to higher vocational education (HBO). If I were to assign every individual the years of education indicated on the figure, then children who finish MBO are considered higher educated, whereas, in practice, they are not.

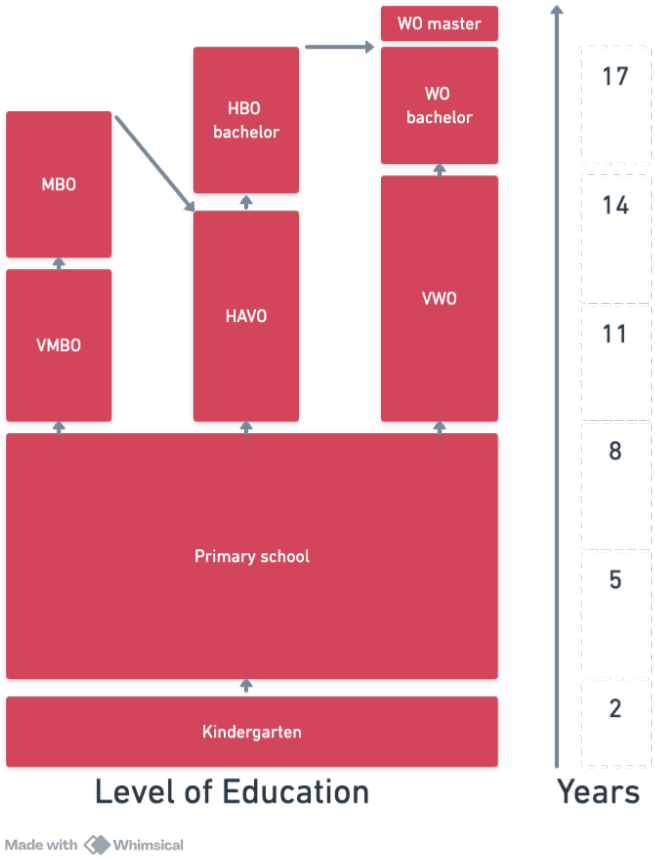



Figure E1: The Dutch Educational System

To overcome this problem, I assign the years of education based on the minimal number of years it can take for students to be eligible for the same follow-up education. For example, individuals with an MBO degree are assigned 13 years of education, which is the same as

children with a HAVO degree. Based on these rules, the conversion table is as follows:

Table E1: Conversion Table of Educational Levels

| Level (Dutch) | Level (International) | Years of Education |
|---|---|---|
| Kindergarten | Kindergarten | 2 |
| Primary school | Primary school | 8 |
| VMBO (all types) | Preparatory vocational education | 11 |
| Practical education | Lower vocational education | 11 |
| MBO 1 | Vocational education (short track) | 11 |
| MBO 2, MBO 3 | Vocational education (medium track) | 12 |
| MBO4 | Vocational education (long track) | 13 |
| HAVO | Preparatory applied science education | 13 |
| VWO | Preparatory academic education | 14 |
| HBO associate | Higher education (fast-track, applied sciences) | 15 |
| HBO bachelor | Higher education (undergraduate, applied sciences) | 17 |
| WO bachelor | Higher education (undergraduate, academic track) | 17 |
| WO master | Higher education (graduate, academic track) | 18 |
| Doctorate | Doctorate | 22 |