TI 2025-010/V
Tinbergen Institute Discussion Paper

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

Sander de Vries[1]

1 Vrije Universiteit Amsterdam

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

Sander de Vries*

February 12, 2025

Abstract

This paper provides new insights on the importance of family background by linking 1.7 million Dutch children's incomes to an exceptionally rich set of family characteristics — including income, wealth, education, occupation, crime, and health. Using a machine learning approach, I show that conventional analyses using parental income only considerably underestimate intergenerational dependence. This underestimation is concentrated at the extremes of the child income distribution, where families are often (dis)advantaged across multiple dimensions. Gender differences in intergenerational dependence are minimal, despite allowing for complex gender-specific patterns. A comparison with adoptees highlights the role of pre-birth factors in driving intergenerational transmission.

Keywords: intergenerational mobility, inequality of opportunity

JEL Codes: I24, J12, 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, and Colchester. 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

Researchers have long been interested in the importance of family background for children's economic success. Motivated by the pioneering work of Becker and Tomes (1979), economists have focused heavily on associations between children's and parents' incomes.[1] These associations are commonly used to quantify intergenerational dependence and compare it between countries, regions, or over time (Blanden (2013), Chetty et al. (2014), Davis and Mazumder (2024)). While this approach generates great insights into how income transmits from parents to children, it leaves open important questions about the importance of the broader family background for children's income.

To get further insights, researchers frequently estimate sibling correlations (Solon (1999)). By construction, these correlations capture all the influences siblings share—not just parental income. However, this approach has three key limitations. First, a substantial share of shared influences may stem from factors unrelated to family background, such as community effects, common shocks, or spillovers (Collado et al. (2023)). Second, it does not identify which dimensions of family background matter most, leaving the underlying drivers unexplained.[2] Third, sibling correlations offer little guidance on where in the income distribution family background exerts the greatest influence, though such knowledge is crucial for equity assessments and targeted policy (Hufe et al. (2022)).

This paper studies the importance of the broader family background while addressing the limitations above. I do so as follows. I first link over 1.7 million Dutch children's adult incomes to detailed information on their fathers' *and* mothers' income, assets, debt, occupation, education, criminal behavior, health, family structure, and various outcomes for aunts and uncles. To my knowledge, this is the most comprehensive project to date linking multiple family background characteristics to children's income with administrative data. I then use these data in combination with a flexible machine learning model to predict child income,

[1] See Solon (1999), Black and Devereux (2011), and Mogstad and Torsvik (2023) for reviews.

[2] Recent papers try to overcome this by integrating parental income, neighborhoods, and schools explicitly into this framework (Bingley and Cappellari (2019), Bingley et al. (2021)).

and compare the explanatory power (R^2) to that of a simpler model using parental income alone. This comparison shows how much conventional estimates of intergenerational dependence increase when a broad range of *observable* family background factors is considered.[3] Beyond these aggregate measures, I (i) present the full distribution of children's expected income ranks—enabling precise identification of the least and most advantaged families, (ii) highlight the family characteristics most strongly associated with income, and (iii) extend the analysis to children's education and criminal behavior.

Incorporating all family background information increases the R^2 from 10.5 percent with parental income alone to 16.6 percent, marking a 60 percent increase. The comprehensive model is particularly more effective at identifying highly (dis)advantaged families. For instance, the 1 percent of children with the lowest expected incomes based on parental income only have an average income rank of 31. With the comprehensive model, this drops to 19. These children face multiple disadvantages: their parents are often young, separated, have low income and wealth, limited education, poor health, and criminal records, with similar disadvantages common among their aunts and uncles. The strongest predictors are parental and extended family income and wealth, highlighting their central role in measuring intergenerational dependence. Additionally, the increase in explanatory power is even larger for children's completed education (102%) and sons' criminal behavior (158%).

Another key open question is whether specific family backgrounds affect sons and daughters differently. I provide novel insights by training separate predictive models for sons and daughters, allowing for unexplored and potentially complex, gender-specific effects of family background characteristics. The results reveal only minor differences: the overall explanatory power for predicting income is similar for boys and girls, and for predicting education, it is slightly higher for girls. Moreover, predictions for sons and daughters are almost perfectly correlated, suggesting that the key family characteristics driving these predictions are the same. This conclusion is reinforced by the family background variables' inability to

[3] I focus on the R^2 because it is easily comparable to conventional mobility measures. Section 2 discusses how this approach relates to sibling correlations or standard regression estimates.

meaningfully predict income or education differences between brothers and sisters.

I present two extensions to distinguish between broad mechanisms driving intergenerational dependence. First, I show that predicted income differences remain accurate even among individuals from the same neighborhood, migrant group, or extended family, suggesting that such broader community factors cannot explain intergenerational dependence well. I then differentiate between pre-birth and post-birth factors by comparing international adoptees raised in families with different levels of advantage. The results indicate that being raised from infancy in a family that is associated with a 1 rank higher income for own-birth children increases the income rank of adoptees by only 0.3. This provides strong evidence that a substantial share of intergenerational transmission is rooted in pre-birth factors.

This paper makes three contributions. First, it offers new insights into the importance of family background for children's long-run income. As discussed above, prior work often relies on sibling correlations that are constrained by their dependence on unobservable factors.[4] While existing studies do link many of the family background variables studied in this paper to child outcomes, they typically analyze one variable in isolation and align it with the outcome of the child.[5] As a result, we know little about the relative importance of each background dimension or its relevance for children's long-run income. This paper addresses that gap by analyzing these family background characteristics jointly and relating them to children's long-run income.

Second, this paper contributes to the small but growing strand of literature on intergenerational dependence that incorporates multiple family characteristics.[6] Most closely related are recent papers using machine learning to predict children's income (Blundell and Risa

[4]This limitation also applies to name-based estimators of intergenerational dependence (Santavirta and Stuhler (2024)). As for sibling correlations, there are numerous unobservable factors beyond family background that may contribute to the similarities among individuals with the same names.

[5]For instance, Black and Devereux (2011) review studies on the transmission of wealth, jobs, occupations, welfare receipt, and health.

[6]Recent 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 income mobility based on extended family income, and Eshaghnia et al. (2022), who measure mobility using expected lifetime income, which is based on multiple parental characteristics.

(2019), Brunori et al. (2023), Brunori et al. (2024), Chang et al. (2025)).[7] 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. Moreover, while previous studies focus on aggregated summary statistics, this paper provides a substantially more detailed analysis by reporting the full distribution of expected incomes alongside the corresponding family background characteristics.

Third, this paper sheds new light on gender differences in intergenerational dependence. Most previous work focuses on son and daughter differences using pooled parental income or paternal income (Chadwick and Solon (2002), Olivetti and Paserman (2015), Davis and Mazumder (2024)). Remarkably few studies use information about fathers or mothers separately.[8] This is the first paper to consider gender differences in intergenerational dependence that uses multiple family background measures and which allows for highly complex interaction effects between paternal or maternal characteristics and sons' or daughters' outcomes. In addition to predicting income and education *levels* separately for boys and girls, I further extend the analysis by predicting brother-sister *differences*. This allows for an accurate analysis of how gender gaps in income and education are related to family background.[9]

This paper proceeds as follows. Section 2 presents a theoretical framework that links measures of intergenerational dependence based on regression estimates, sibling correlations, or predictive models. Section 3 presents the data and Section 4 discusses how this is used for training the machine learning models. Section 5 presents the main results. Sections 6 and 7 analyze gender differences and mechanisms, respectively. Section 8 concludes.

[7] Brunori et al. (2023) and Brunori et al. (2024) come from a related literature that quantifies inequality of opportunity. In section 2, I discuss how this literature relates to the approach in this paper.

[8] Notable recent works that address this gap are Brandén et al. (Forthcoming) and Ahrsjö et al. (2023), who study gender-specific trends in intergenerational mobility in Scandinavian countries, and Althoff et al. (2024), who study historical trends in the US between 1850-1940 using multiple parental inputs, including separate measures of maternal and paternal human capital.

[9] Some studies focus specifically on gender differences in the most disadvantaged families (Bertrand and Pan (2013), Chetty et al. (2016) Brenøe and Lundberg (2018), Autor et al. (2019), Lei and Lundberg (2020), Autor et al. (2023)). I report results for different types of (dis)advantaged families in section 6.

2 Theoretical Framework

This section presents a simple framework linking the approach in this paper to intergenerational mobility estimates, sibling correlations, and inequality of opportunity estimates. The analysis is at the population level. Model estimation, evaluation, and inference are discussed in section 4.

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.[10] Consider the following two conditional expectations function decompositions of Y_{sf}:[11]

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

[10] 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.

[11] 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.

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 ν_{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. I call the fraction of the variance in $f(\mathbf{X}_f, \mathbf{Z}_f)$ that is explained by $g(\mathbf{X}_f)$ the models' *sibling-completeness*:[12]

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

where $R^2_{y|f} = V(f(\mathbf{X}_f, \mathbf{Z}_f))/V(Y_{sf})$ 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 $\mathbf{Z}_f$.[13]

[12] Blundell and Risa (2019) consider another measure of completeness. They assume that a tuned machine learning model with many variables is the complete model and assess how well a linear model with only income performs relative to this model.

[13] 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.

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, β, 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.[14] 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.

3 Data

I use administrative data on the entire population of the Netherlands from Statistics Netherlands.[15] Individual identifiers enable me to join records associated with an individual across a range of government services, such as the personal register, tax statements, enrollments in education, crime incidents, neighborhood residency, and healthcare insurance reimbursements. This section describes the sample selection, outcomes, explanatory variables, and descriptive statistics.

[14] 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) discuss in detail how intergenerational mobility coefficients and inequality of opportunity estimates are related.

[15] The administrative data from Statistics Netherlands is available at a remote-access facility after signing a confidentiality agreement.

Sample. For the main analysis, I consider all children born in the Netherlands between 1980 and 1989. I drop 3.4 percent of children due to missing income records of the children, mostly because of emigration and a small portion due to death. This yields a sample of 1,704,065 children. For the education and crime analyses, I focus on children born between 1985 and 1989. This is due to the unavailability of suitable education or crime records for earlier periods. I exclude 0.5 percent of children from the education sample due to missing records, resulting in a sample of 908,876 children. The crime analysis focuses exclusively on boys, resulting in a sample of 463,625 children.

Children's household income. The main outcome in this paper is a child's long-run household income. 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 intergenerational mobility studies (Chadwick and Solon (2002)). The income register records the gross household income extracted from (joint) tax statements spanning the period between 2003 and 2023. Household income encompasses 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.[16][17] I measure income in 2024 euros, adjusting for inflation using the consumer price index.

I use the income data to construct a proxy for children's lifetime household income. A well-known challenge is that snapshots of an individual's income are prone to measurement error due to transitory income shocks (Mazumder (2005)) and life-cycle bias arising from heterogeneous age-income profiles (Haider and Solon (2006)). To mitigate these issues, I

[16]Income insurance payments concern benefits from social insurance, national insurance and private insurance related to unemployment, illness, disability or retirement. Social security benefits concern government-sponsored transfers such as welfare benefits or veteran pension payments. Conditional transfers are transfers tied to specific payments, such as rental or study allowances. Receiving income transfers consist of transfers between households such as alimony received from the ex-spouse.

[17]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.

average each individual's household income over all available years starting from age 30.[18] Because the oldest children were born in 1980, income is measured at most up to age 43. Overall, 96 percent of children have at least five years of income data contributing to their average household income, with a mean of nine income observations per child.

I then define children's income ranks based on their positions in the distribution of long-run household income in their respective cohorts. I focus on ranks due to their low attenuation bias, stability over the life cycle, and ease of comparability with other intergenerational mobility research that frequently adopts rank-based metrics (Chetty et al. (2014), Nybom and Stuhler (2017)). For robustness, I also provide results using alternative specifications, including averaging incomes over different time spans or ages, as well as income levels or personal income ranks instead of household income.

Children's education and crime. The education registers contain individuals' highest attained education until 2022. I use these data to construct a years-of-education variable according to the conversion table in Appendix D.

The crime register 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. 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 and because these provide a unique manifestation of lower-level acquisition of non-cognitive skills in my data.[19]

[18] I exclude years in which household income falls below €1,000, as these often correspond to individuals with significant wealth but low reported income in that year.

[19] Violent crime includes the following categories from the Dutch penal codes: theft with violence, robbery, assault, public violence, violence against a civil servant, stalking, crime against life, kidnapping/deprivation of liberty, human trafficking, threat, sexual assault, rape, and other violent crime.

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. The parental income rank is then based on the position within the parental income distribution of all children in the analysis sample.

Other explanatory variables. The other family background variables are motivated by prior research demonstrating that, beyond income, parents' education, wealth, health, occupation, criminal behavior, family structure, and similar variables of aunts and uncles are all predictive of child outcomes. These dimensions reflect key aspects of 'socioeconomic status', frequently studied by economists and sociologists. Table 1 describes how all 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, because Statistics Netherlands began systematically storing administrative data primarily in the 2000s, 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.[20]

[20]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

Second, as shown in Table B1, despite the extensive coverage of variables, some gaps persist. For instance, education records for the parents' generation are incomplete, as systematic recording began in the 1980s. Additionally, data on fathers are missing for 3 percent of children, resulting in the omission of all related paternal outcomes in those cases. 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.

Table 1: Explanatory Variables

Income	Average values of the following income variables after 2003 and up to age 60: household income, personal income, and personal earnings. I also compute the most important sources of personal income over this period (in 11 categories), and the share of household income due to transfers.
Wealth	Average value of the following types of household assets and debt between 2006 and 2011: 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) between 2006 and 2009.
Education	Highest level of completed education.
Health	Average healthcare costs between 2009 and 2011 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 between 2005 and 2010.
Family structure	Parents' family size, age-at-first-birth, birth order, household type†, father or mother presence†, parental death†, child family size, and whether the father or the mother are identified in the child-parent register.
Extended family outcomes	Average years of education, household income rank, wealth rank, total health 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).
†: Household type consists of three categories: registered partner, non-registered partner, or single. Father (mother) presence is an indicator of whether the father (mother) is registered in the same household as the child. Household type, father/mother presence, and parental death are all measured at age 15 of the first child.

the age of the children at which these resources are measured.

4 Model training and evaluation

The objective is to train a predictive model, $\hat{g}$, that accurately predicts the conditional expectation function g (see equation 2). A key challenge is that the true functional form of g is unknown. Variables may enter g 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 Friedman (2001).[21] Single decision trees partition the covariate space into regions with similar outcome values, predicting the average value for new observations within the same region. Gradient-boosted trees refine predictions by employing multiple trees, where each successive tree is trained on the residuals of prior trees. This iterative process allows gradient-boosted trees to effectively handle highly non-linear data-generating processes with complex interactions between variables. Their complexity depends on several parameters, including the maximum number of splits per tree, the minimum gain required for a split, the total number of trees, and the learning rate.

Model training and evaluation proceed as follows. 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). The training data is used to optimize parameters and train the model, while the test data is reserved for evaluation, ensuring the model has not seen the observations it predicts. Specifically, 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 $R^2_{y|\hat{g}}$. Generally, all results in this paper that rely on predictions are based on observations from the test data.

Sampling variability affects estimates of explanatory power in two distinct steps. First, the model is trained on a specific draw of the training data. Second, the model's explanatory power is evaluated using a specific draw of the test data. Both steps will be numerically

[21] I also experimented with other machine learning methods, which produced similar or inferior performance.

different for different draws and sizes of the data, and so are subject to uncertainty. To gauge the uncertainty arising from the first step, I test the sensitivity of prediction error to sample size reductions. If, even with less data, the prediction error remains constant, then this suggests that the uncertainty from the first step is low. Taking the tuned model $\hat{g}$ as given, I gauge uncertainty from the second step by computing confidence intervals for $R^2_{y|\hat{g}}$ using a bootstrap.[22]

5 Main Results

5.1 Intergenerational Income Mobility in the Netherlands

This section provides a baseline analysis of intergenerational income mobility in the Netherlands and compares it to similar estimates from other countries.

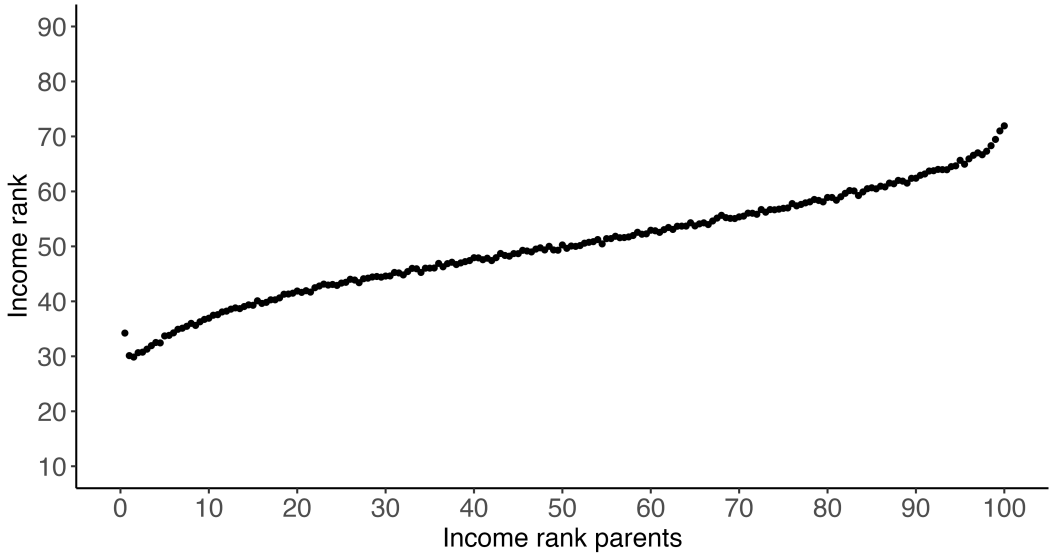
Figure 1 presents a scatter plot of children's income ranks relative to their parents' income ranks. The X-axis is divided into 200 bins, each representing half a percentile and containing roughly 8,500 children. The dots correspond to the mean household income rank of children given their parents' household income rank. Child income increases linearly between the 10th and the 90th income ranks but increases steeply at the tails of the parental income distribution.[23] Such an inverse S shape is commonly found in other countries. An OLS regression yields a slope coefficient of 0.32, indicating that a one-rank increase in parental income corresponds to a 0.32-rank increase in children's income on average.

The rank-rank correlation of 0.32 positions the Netherlands among the developed 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),

[22] Ideally, the bootstrap is applied to both steps simultaneously, such that for each draw of the data b, a new model $\hat{g}_b$ is trained *and* evaluated. However, as tuning the machine learning models is time-consuming, this is computationally infeasible. As such, I analyze uncertainty from the first step separately.

[23] As noted before by Van Elk et al. (2024), there is some measurement error at the very bottom of the parental income distribution. Some of these parents have extremely low income but high wealth. Removing the bottom 0.5 percent of the sample does not affect the estimates much.

Figure 1: 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 $N = 1,703,392$ 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.

similar to France, Italy, and the UK (~ 0.30), and lower than in the US (0.36).[24] Despite the Netherlands' reputation for relatively low-income inequality and affordable, high-quality education, intergenerational mobility appears surprisingly low.[25]

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 estimates are robust to measurement error and lifecycle bias. Consequently, additional explanatory power gained from incorporating more variables is unlikely due to these variables merely correcting for measurement error in the parental income variable.

[24] 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).

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

5.2 Including Detailed Parental Information

The previous section demonstrates a substantial degree of intergenerational dependence based on parental income only. This section explores by how much incorporating the additional observable dimensions detailed in Section 3 increases estimates of intergenerational dependence. It also identifies where along the income distribution these increases are most pronounced and highlights key family characteristics linked to low or high child income.

To quantify the overall increase in intergenerational dependence, 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. 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 4.

Adding all information about the parents reveals substantially stronger intergenerational dependence. The comprehensive model achieves an explanatory power of 16.6 percent, marking a 60 percent increase compared to the income-only model. To put this into perspective, increasing the rank-rank correlation from 0.32 to 0.41 results in the same increase in R^2.[26] While this may seem modest, it is significant, 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 using all available income data ($R^2 = 10.5\%$) versus one year of income data ($R^2 = 7.6\%$) in a rank-rank regression.[27] This source of measurement error has received considerable attention in the literature (Mazumder (2005), Nybom and Stuhler (2017)).

While the explanatory power is already interesting by itself, it does not reveal how family

[26] I use here that in a rank-rank regression, $R^2 = \beta^2$ (i.e. $0.41^2 - 0.32^2 = 0.166 - 0.105 = 0.061$).
[27] See Table A2 columns 1 and 9.

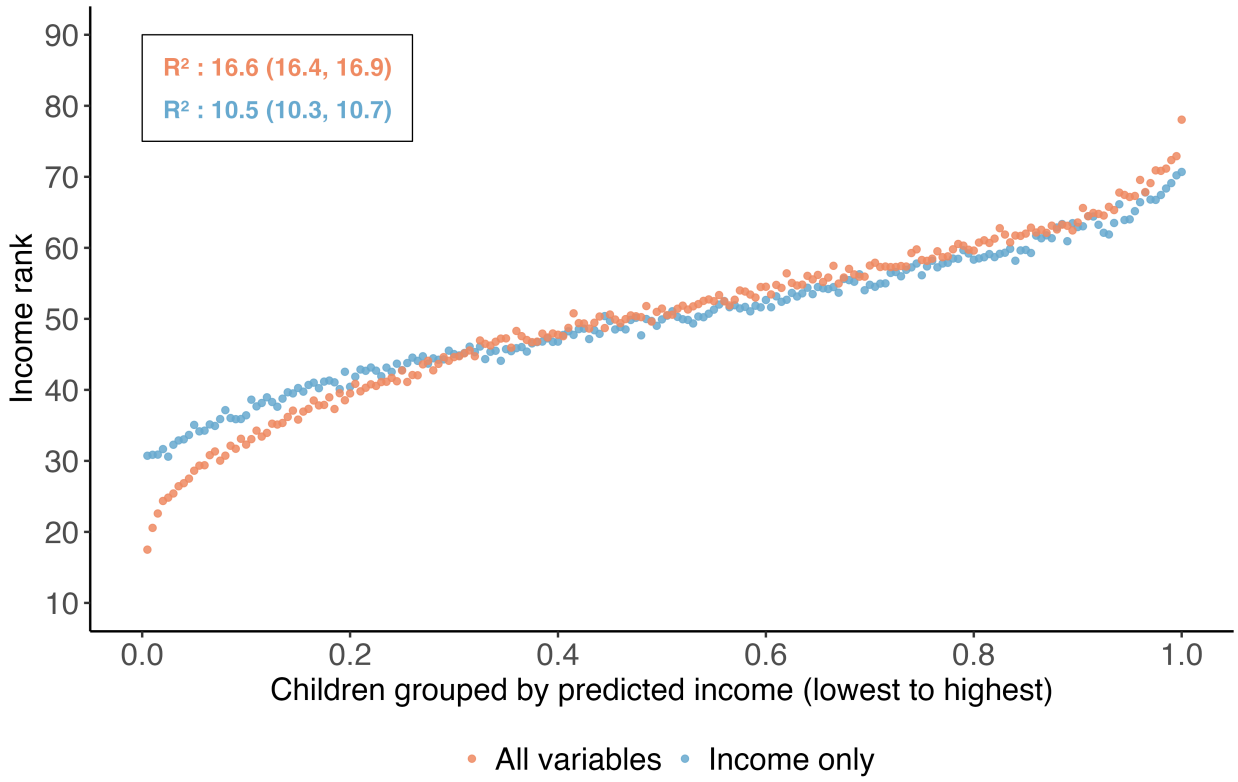
background effects vary. For instance, are there family types where children have exceptionally high or low income? To explore this, Figure 2 provides a binscatter plot of children's income ranks, sorted from lowest to highest predicted income. Specifically, the X-axis divides the test dataset into 200 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.[28]

Figure 2 shows that the comprehensive model is particularly more effective at identifying highly (dis)advantaged families. 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. Even within this top 1 percent there are striking differences: children in the top 0.5 percent reach an average rank of 78, five ranks higher than the next 0.5 percent.

What distinguishes children at the bottom and top of the predicted income distribution? Table 2 highlights some family background characteristics across the predicted income distribution, focusing on the most advantaged and disadvantaged children. Each column bins children from a distinct group: the first column includes the 0.5 percent of children with the lowest predicted incomes, the second column covers the next 0.5-1 percent of children with the lowest predicted incomes, and so forth. 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 family background characteristics.

[28]The blue graph in Figure 2 closely resembles the black graph in Figure 1. There are two differences: it shows only observations in the test data, and children are ranked by predicted income rather than parental income. These rankings differ slightly due to small non-monotonicities in the relationship between parental income and mean child income, as shown in Figure 1.

Figure 2: Predicting Child Income with Detailed Parental Information



Notes: this figure presents binscatter plots of income ranks for 340,813 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 1,704,065 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								
	0-0.005	0.005-0.01	0.01-0.05	0.05-0.1	0.1-0.9	0.9-0.95	0.95-0.99	0.99-0.995	0.995-1
Child income rank	17.7	20.5	25.8	31.2	50.5	65.7	69.9	72.8	78.1
Parental income rank	5.9	8.0	11.5	16.1	49.3	87.8	93.2	97	98.6
Parental wealth rank	13.5	14.9	15.6	17.2	50.6	75.4	81.4	88.0	90.7
Max. education parents	8.1	8.6	9.4	9.9	13.1	16.1	16.7	17.1	17.4
Health costs parents	4,532	4,231	3,843	3,719	2,571	1,844	1,816	1,818	1,782
Crime father	0.56	0.42	0.30	0.19	0.05	0.02	0.03	0.03	0.03
Extended family income	17.2	20.4	25.1	30.4	49.3	64.4	69.3	74.9	79.3
Extended family wealth	19.4	21.6	24.7	29.1	51.0	64.8	68.8	73.5	76.7
Father presence	0.32	0.35	0.48	0.63	0.88	0.97	0.98	0.99	0.98
Age at first birth mother	22.1	22.9	24.0	25.2	27.1	28.4	28.7	29.0	28.9
N	1,704	1,704	13,632	17,041	272,650	17,041	13,632	1,704	1,705

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.

Table 2 shows that children at the extremes face multiple (dis)advantages. The first four columns show that the most disadvantaged children have parents with low income and wealth and who are often minimally educated, have high health expenditures, and are often suspected of crimes. Their aunts and uncles also have low income and wealth, and their parents are often young and separated. In contrast, the family background characteristics of the most advantaged children are the polar opposites of those of the disadvantaged children.

Although the variables in Table 2 are all correlated with child income, they are not equally good predictors. In Appendix C, I present 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 intergenerational dependence.

5.3 Additional results

Sibling completeness. While models with income only underestimate intergenerational dependence, sibling correlations can overestimate it by capturing all influences shared between siblings. The sibling correlation in income is 0.31.[29] This implies that the sibling completeness of the comprehensive model is about 50 percent (0.16/0.31, see equation 4). The remaining half of the sibling correlation may be explained by other shared factors, such as community influences, shocks, or spillovers, that are uncorrelated with the included variables.

The previous subsection highlights two advantages of the prediction approach over sibling correlations. First, while both provide aggregate measures of explanatory power, the predictive model is more transparent due to its reliance on observable inputs. Second, the prediction approach enables visualization of the full distribution of expected incomes, revealing strong patterns of intergenerational dependence at the tails. The sibling correlation approach relies on many imprecisely estimated family fixed effects, making it unsuitable for such a granular analysis.

Functional form. How much of the improvement in explanatory power can be attributed to additional variables versus using a flexible machine-learning model? A straightforward OLS model, which includes all variables linearly, achieves an explanatory power of 15.3 percent.[30] This is close to the explanatory power of the comprehensive model, suggesting that incorporating a broader range of information is more critical than allowing for complex interactions and non-linearities.

Robustness. Table B2 shows that the explanatory power estimates are consistently high across different sample sizes, even when using only 1 percent of the core sample (approximately 17,000 observations). This demonstrates that while administrative data provide comprehensive coverage, similar analyses can be effectively conducted with smaller datasets. Table B3 varies the number of years and ages at which child income is measured. The results

[29] This is estimated by the adjusted R^2 of a regression of income on family fixed effects and birth year with the core sample.

[30] The 95% confidence interval is (0.152, 0.154). Coefficient estimates are available upon request.

from this analysis mimic that of the robustness of the rank-rank correlation in Appendix A. 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 in the early 30s, but stabilizes after age 34. This indicates that the influence of attenuation or life-cycle bias is likely minimal.

Predicting income levels. The heightened intergenerational dependence shown in Figure 2 raises concerns for societies that are averse to disparities in expected income driven by family background. To further illustrate these equity implications, Figure B1 presents results analogous to those in Figure 2, but with models predicting income levels instead of ranks. This figure highlights the scale of expected income disparities. For instance, the gap between the top 1 percent and bottom 1 percent of expected incomes is about €120,000 — equal to 1.2 median incomes or 1.9 standard deviations.

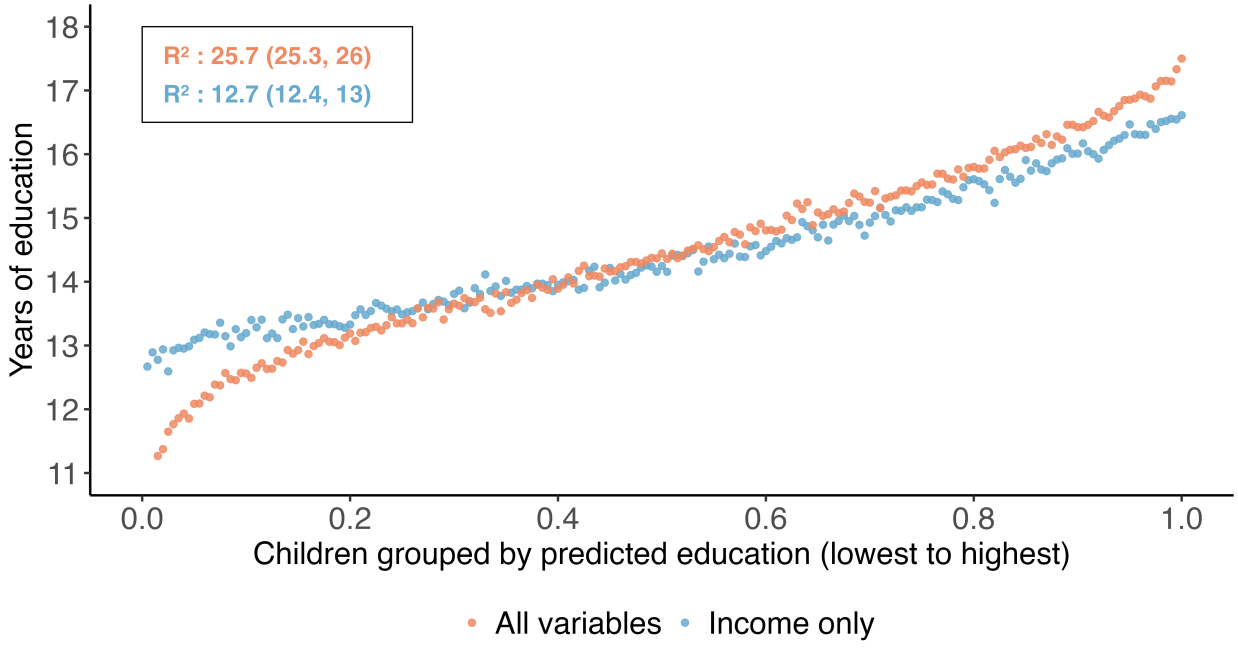
5.4 Predicting Education and Crime

This section reports results for children's completed education and criminal behavior. As violent crimes are predominantly committed by men, I focus on men's criminal behavior only, but in Appendix B2, I report results for women's criminal behavior too.

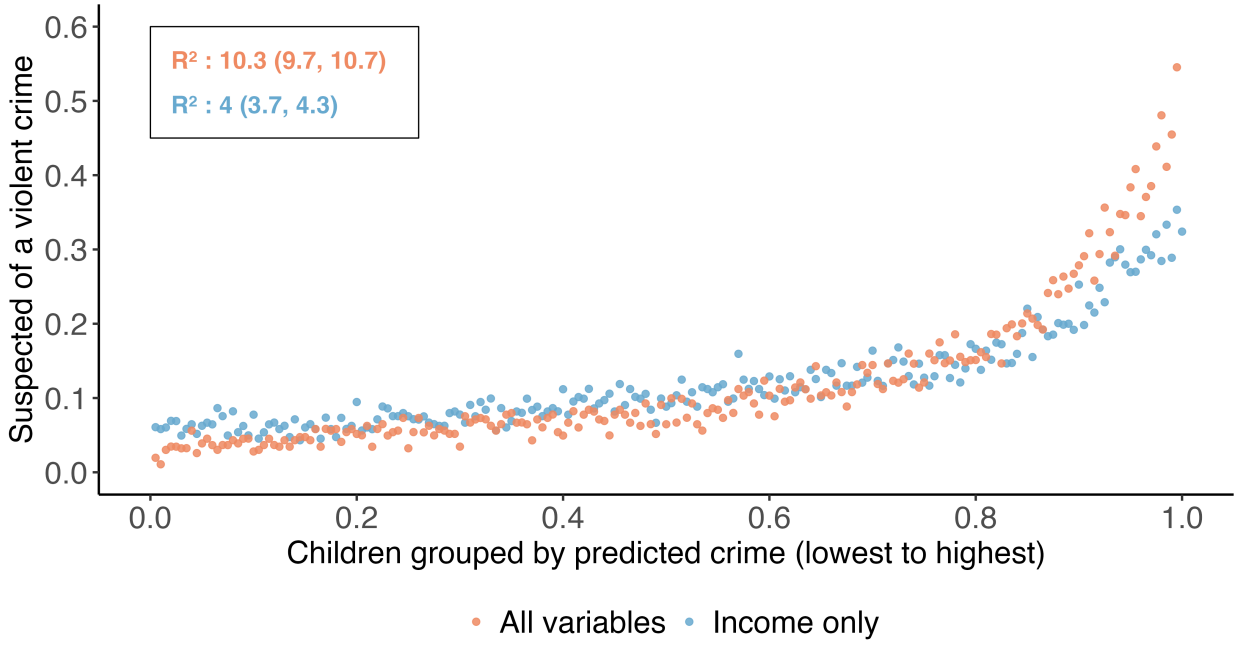
Figure 3 (a) presents the results for education. The explanatory power of the income-only model is 12.7 percent.[31] Incorporating all explanatory variables significantly boosts explanatory power for education, doubling it to 25.7 percent. This increase in explanatory power is considerably larger than for income. The graphs reveal strong differences in children's education by family background. For example, children with the 5% lowest predicted education levels have on average less than 12 years of education, frequently dropping out without essential qualifications, whereas children with the 5% highest predicted education levels have on average 17.1 years of education, corresponding to an undergraduate degree.

[31] 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.

Figure 3: 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 2. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and are reported in brackets.

Figure 3 (b) shows a similarly large increase in explanatory power for crime, from 4 percent for the model that incorporates income only to 10.3 percent for the comprehensive model.[32] The results indicate that violent crime is highly concentrated in certain families. A simple calculation shows that the 20 percent of boys with the highest crime risk in Figure 3 (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.

6 Gender Differences

This section analyzes gender differences in intergenerational dependence. To do so, I estimate separate predictive models for sons and daughters, using gender-specific training and test datasets. Training the predictive models separately for each gender allows for the possibility that different characteristics of fathers or mothers influence sons and daughters in unique ways. The results are reported in Figure 4, illustrating distinct patterns for each gender.

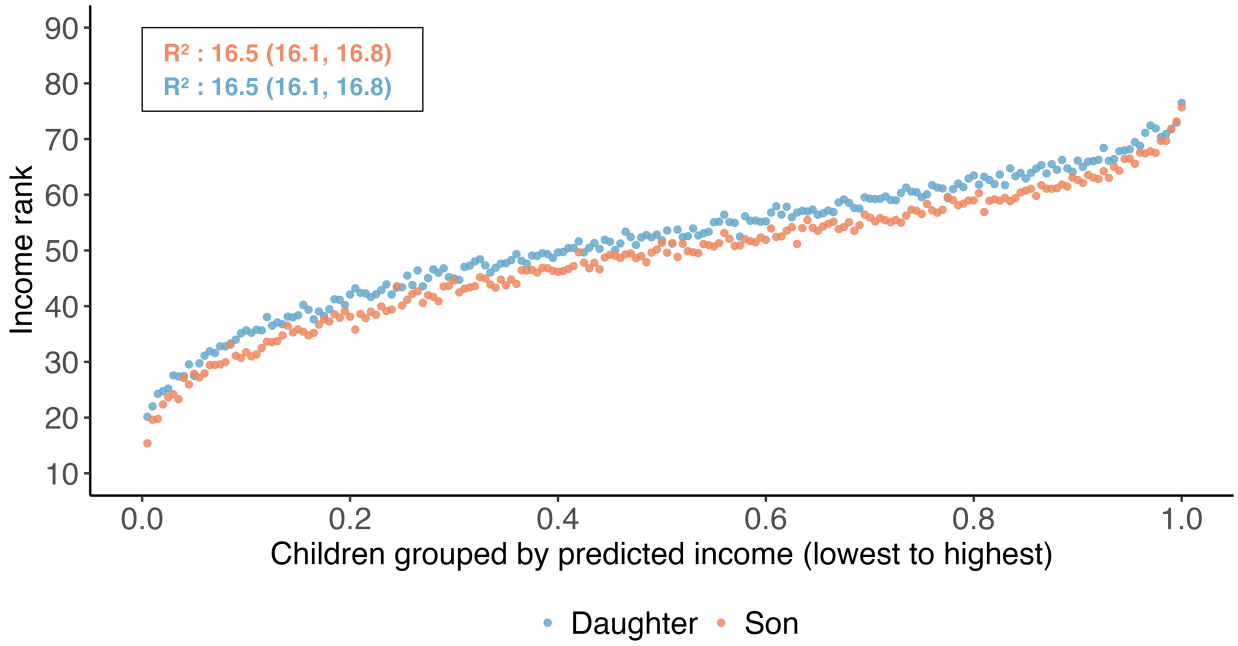
Figure 4 (a) shows that gender differences in intergenerational income dependence are minimal. The models have practically identical explanatory power across genders, with R^2 values of 16.5% for both sons and daughters. The nearly parallel prediction lines for both genders also suggest that daughters' and sons' incomes are similarly influenced by their family background.[33]

Figure 4 (b) presents results for education, showing a slightly higher explanatory power for daughters (27.7 percent) than for sons (24.6 percent). Figure 4 (b) shows a consistent

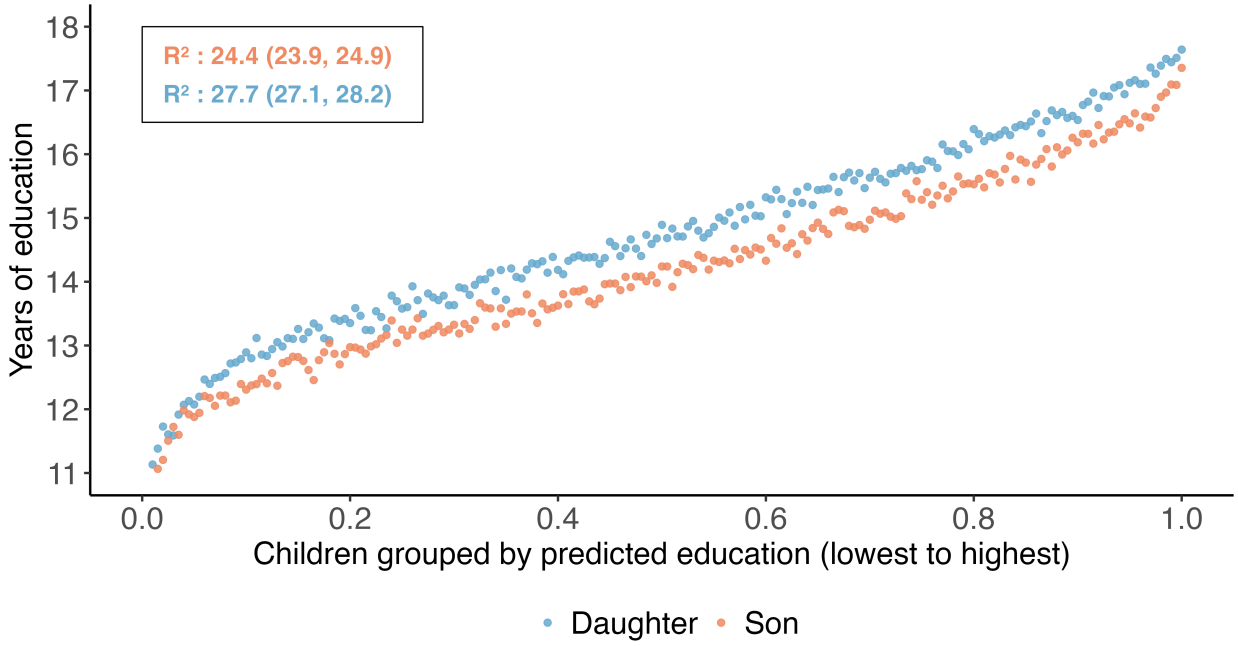
[32] The family background characteristics across the predicted education and crime distributions show very similar patterns to those seen in income predictions in Table 2. Results are available upon request.

[33] Daughters have slightly higher household income than sons across the predicted income distribution. This pattern is due to men being single more often and for longer periods during their 30s, resulting in lower household income. To abstract from such issues, I also perform the analysis using personal income ranks as outcomes in Figure B3. With personal income, a clear gender gap favoring men emerges. This gap is relatively constant across the predicted income distribution and the overall explanatory power is a bit higher for daughters (17.7 percent) than for boys (16.6 percent).

Figure 4: Predicting Children's Income and Education by Gender



(a) Income



(b) Education

Notes: the figures above display scatter plots of sons' and daughters' income ranks and years of education, ordered by their predicted value. Predictions are generated using the same predictive model and explanatory variables as in Section 5, now applied separately to each gender. The construction of the graphs follows the same steps as in Figure 2, now separately for each gender. Panel (a) reports the results for 173,775 sons and 167,039 daughters in the test datasets. Panel (b) reports the results for 92,229 sons and 88,601 daughters in the test datasets. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and reported in brackets.

gender gap favoring daughters, which narrows at the extremes: the least advantaged daughters have similar education levels as their male peers, and the most advantaged daughters achieve education levels comparable to the most advantaged sons. This pattern may partly reflect floor and ceiling effects, as less than 11 years of education corresponds to dropping out of school, while more than 17 years indicates completing a master's degree. Overall, these results suggest that daughters' education is somewhat more sensitive to their family background, although the differences are small and only pronounced at the tails of the education distribution.

The figures above indicate that the level of intergenerational dependence is relatively similar for boys and girls. However, the underlying mechanisms that generate these levels may differ. This is the case when certain family environments favor sons over daughters, and vice versa. To investigate this, I first generate separate income predictions for boys and girls for each family and compute their correlation.[34] This correlation is 97 percent, indicating that the family background characteristics that predict boys' income well, also predict girls' income well. The predicted education levels are equally highly correlated.

To explore this further, I focus on families with at least one son and daughter, and compute the difference between the average income of the boys and the girls. If certain family backgrounds systematically favor one gender, these gaps should be predictable. However, consistent with minimal gender differences, the explanatory power for predicting them is just 0.13 percent. Table B4 reports characteristics of families with low and high predicted income gaps, confirming that even the largest predicted income differences remain small. The only notable trend is that brother-sister gaps mirror father-mother differences: sisters have higher income in families where mothers earn relatively more, and lower when fathers earn relatively more. However, these gaps remain modest relative to overall income variation. Similar patterns emerge for education.

Some previous papers have particularly focused on gender differences in long-run out-

[34] I predict two incomes per family, regardless of the number or gender of the children.

comes among the most disadvantaged families. While Chetty et al. (2016) find that disadvantaged family environments disproportionally harm boys' long-term outcomes, Brenøe and Lundberg (2018) and Lei and Lundberg (2020) find no such effects in Danish or U.S. data. This analysis incorporates significantly more detailed family background information than previous studies, allowing for a more precise identification of the most disadvantaged sons and daughters. If boys' long-run outcomes were more sensitive to highly disadvantaged backgrounds, this should be apparent in Figure 4 as a larger gender gap at the lower end of the predicted income and education distributions, or in the prediction exercise of brother-sister gaps by larger predicted gaps in disadvantaged families. However, neither analysis provides evidence supporting this pattern

7 Mechanisms

7.1 Neighborhoods, Migration Background, and Extended Family

Family background is strongly correlated with the neighborhoods in which children live, the type of extended family they grow up in, and their migration background. This raises an important question: do the disparities in Figure 2 stem from these broader factors, or do they reflect differences in parental inputs?

To better understand this, I evaluate the models' prediction accuracy for children who come from the same neighborhood, extended family, or have a similar migration background. Specifically, I regress children's income on their predicted values and group fixed effects. These group fixed effects ensure that only individuals from the same neighborhoods, migration background, or extended family are compared. Without these fixed effects, the coefficient of the predicted values is equal to 1 by construction. If this coefficient decreases after adding fixed effects, it indicates that group-specific factors explain some disparities captured by the family background variables.

Table 3 presents the results of these regressions. Column 1 includes neighborhood fixed

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

	Income rank (y)				
	(1)	(2)	(3)	(4)	(5)
Predicted income rank ($\hat{y}$)	0.947	0.907	0.872	0.275	0.863
	(0.005)	(0.015)	(0.014)	(0.026)	(0.047)
Fixed Effects					
Neighborhood	x				
Migration background		x			
Extended family			x		
N	333,930	51,138	523,280	4,938	3,804
Sample	All	Second generation migrants	Extended family sample	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 2. The samples in columns 1 to 5 correspond to the following children from the test data: (1) children with an available neighborhood identifier, (2) second-generation migrants, (3) children with an (identified) extended family, (4) international adoptees, and (5) own-birth children from families with at least one adopted child. Standard errors, shown in parentheses, are clustered at the fixed-effect level in column 1 to 3.

effects, corresponding to the neighborhood where children are registered at age 15.[35] The coefficient in column 1 is 0.948, indicating that *within* neighborhoods, a 1 rank increase in predicted income is associated with a 0.95 rank increase in predicted income. This shows that also within neighborhoods, differences in family background are still strongly associated with differences in children's income.[36]

Column 2 focuses on second-generation migrants, corresponding to 15% of the sample, and includes migration background fixed effects. Each fixed effect corresponds to the region of origin of the father, mother, and grandparents whenever available. As such, I restrict the comparisons to individuals whose fathers, mothers, and grandparents come from the same region. I consider eight regions of origin: Netherlands, Morocco, Turkey, Suriname, Dutch Antilles, Western Europe, Eastern Europe, and others. The coefficient in column 2

[35]The neighborhood code is based on the most granular level of Statistics Netherlands' neighborhood classifications (in Dutch: "buurt"). Neighborhoods are measured at a very granular level with mean and median population sizes of about 1500 and 900 individuals, respectively. I have neighborhood identifiers for 95% of the children.

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

is also close to one, indicating that differences in migration background also can not explain disparities by family background.

Arguably the most restrictive comparisons are made with grandparent-fixed effects, shown in column 3.[37] Since cousins often grow up in similar regions, share family traditions, and are genetically related, these fixed effects may absorb many factors related to parents' characteristics. Although the coefficient decreases more than in previous specifications, it remains close to one.

Overall, while neighborhoods, migration backgrounds, and extended families do seem to play a role, they can explain only a small portion of the disparities observed.[38] The most likely hypothesis is therefore that parental inputs are the main drive behind the disparities observed in the main results.

7.2 The Post-birth Environment

This subsection examines the causal effect of being assigned shortly after birth to a family associated with a 1 rank higher predicted income. This analysis provides insight into the extent to which the disparities observed in Figure 2 are driven by post-birth factors, as opposed to pre-birth influences such as genetic transmission and in-utero conditions.

To answer this question, I use a sample of international adoptees who were adopted in infancy. Although the Netherlands lacks a centralized adoption register, Statistics Netherlands developed a method to identify adoptees reliably. They classify a child as an adoptee if the child was born in a country with many known adoptions to the Netherlands and at least one parent was born in the Netherlands. I expanded this method by including only children who arrived in the Netherlands within six months of birth. This produces a sample of 4,938

[37] Since children have two couples of grandparents, some children occur twice and are compared once to cousins from the father's side and once to the cousins from the mother's side. Some children are dropped because they have no identified extended family and some children are used only once because they have only one identified extended family.

[38] Note that this does not imply that policies targeted at neighborhoods, migrants, or extended families are likely ineffective at reducing disparities.

adopted children.[39] 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.[40]

Table 3 column 4 shows the coefficient that is obtained by regressing adoptees' income on their predicted income. The estimate indicates 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 0.28. Assuming no selection bias and generalizability towards the broader population, this estimate would suggest that around 30% of the disparities in Figure 2 are shaped by the post-birth environment. This result underscores the critical role of pre-birth factors in driving the observed level of intergenerational dependence.

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 waiting times.[41] Second, Table B5 reports estimates from various specifications that include controls for gender, age at migration, and fixed effects for the country and year of adoption - all observable characteristics of the child at the time of adoption. These estimates are all close to 0.28, indicating that selective placements based on these observable characteristics

[39] Statistics Netherlands ran a large randomized survey for a subset of all plausible adoptees, finding that 96.9 percent of the respondents confirm that they are adopted. When I enhance the sample restriction to include only children who arrive within six months of birth and who are born between 1980 to 1989, this increases to 97.7 percent ($n = 778$). If a small fraction of children in my sample are own-birth children, then this likely results in a small upward bias in the estimates.

[40] 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 incorporating richer data on adoptive parents and by focusing on children's long-run income. Despite its central role in descriptive intergenerational mobility analyses, this dimension has been overlooked in studies using international adoptees.

[41] 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.

are of limited empirical importance.[42]

I also offer three reasons why external validity concerns may not be overly severe. First, 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. Second, while there are no highly disadvantaged adoptive families, Appendix Table B6 reveals substantial variation in the characteristics of adoptive families, spanning a broad range of the general population. Third, 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.

Finally, some family background characteristics that strongly predict own-birth children's income may be less predictive for adoptees. While retraining the model on the smaller adoptee sample is infeasible, I can isolate the role of parental income—the strongest predictor of own-birth children's income—from other factors. Appendix Table B5 shows that, holding parental income constant, a 1-rank increase in predicted income corresponds to a 0.4-rank rise in adoptees' income. This higher estimate shows that the other family background characteristics that are strongly associated with own-birth children's income, such as parental wealth and extended family income and wealth, are rooted stronger in post-birth factors than parental income.[43] In other words, including broader family background characteristics not only raises estimates of intergenerational dependence but also exposes a greater role for post-birth factors, which are generally seen as more amenable to policy intervention.

[42] Even if selective placements occurred, the estimate underscores the importance of pre-birth factors. Since the adoptees were adopted in infancy, any selection bias must come from correlations between pre-birth factors and parental characteristics. Therefore, even if selection bias drives the estimate to 0.3, its source is precisely the pre-birth factors themselves. It would, however, complicate the interpretation of the estimate's magnitude.

[43] Section 5 highlights these variables as the most important predictors beyond parental income. These results are consistent with Fagereng et al. (2021) and Adermon et al. (2021), who show that parental wealth and extended family education are relatively strong predictors of international adoptees' wealth and GPA.

8 Conclusion

This study demonstrates that incorporating a broad set of family background characteristics enhances the measurement of intergenerational dependence. Combining comprehensive administrative data from the Netherlands with machine learning techniques, I show that including family background characteristics beyond parental 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. As a result, conventional analyses using income only may give the impression that children's outcomes—especially for the most and least advantaged— are less affected by their parents than they are in reality.

This paper also provides new insights into gender differences in intergenerational dependence by examining previously unexplored differential effects of multiple family background characteristics on sons and daughters. Perhaps surprisingly, these gender differences are small. Not only do boys and girls exhibit comparable overall levels of intergenerational dependence, but the underlying family background characteristics contributing to it are also broadly similar between genders. I conclude that gender gaps in income or education are largely unrelated to these observable family factors.

The broader availability of large datasets and advancements in computing power and statistical methods have greatly enriched our understanding of intergenerational dependence. This paper illustrates how a large number of variables can be analyzed jointly using state-of-the-art machine learning methods that do not rely on traditional linearity and homogeneity assumptions. As more comprehensive data becomes accessible, future research can employ this approach to further explore the intricate intergenerational transmission process.

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, 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.

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.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2019. "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes." *American Economic Journal: Applied Economics* 11 (3): 338–381.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2023. "Males at the Tails: How Socioeconomic Status Shapes the Gender Gap." *The Economic Journal* 133 (656): 3136–3152.

Becker, Gary S., and Nigel Tomes. 1979. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *Journal of Political Economy* 87 (6): 1153–1189.

Bertrand, Marianne, and Jessica Pan. 2013. "The Trouble with Boys: Social Influences and the Gender Gap in Disruptive Behavior." *American Economic Journal: Applied Economics* 5 (1): 32–64.

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.

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.

Blanden, Jo. 2013. "Cross-Country Rankings in Intergenerational Mobility: A Comparison of Approaches from Economics and Sociology." *Journal of Economic Surveys* 27 (1): 38–73.

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.

Brenøe, Anne Ardila, and Shelly Lundberg. 2018. "Gender Gaps in the Effects of

Childhood Family Environment: Do They Persist into Adulthood?" *European Economic Review* 109 42–62.

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. 2023. "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.

Chetty, Raj, Nathaniel Hendren, Frina Lin, Jeremy Majerovitz, and Benjamin Scuderi. 2016. "Childhood Environment and Gender Gaps in Adulthood." *American Economic Review* 106 (5): 282–288.

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.

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.

Hufe, Paul, Ravi Kanbur, and Andreas Peichl. 2022. "Measuring Unfair Inequality: Reconciling Equality of Opportunity and Freedom from Poverty." *The Review of Economic Studies* 89 (6): 3345–3380.

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

Lei, Ziteng, and Shelly Lundberg. 2020. "Vulnerable Boys: Short-term and Long-term Gender Differences in the Impacts of Adolescent Disadvantage.." *Journal of Economic Behavior & Organization* 178 424–448.

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.

Olivetti, Claudia, and M. Daniele Paserman. 2015. "In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850-1940." *American Economic Review* 105 (8): 2695–2724.

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.

Santavirta, Torsten, and Jan Stuhler. 2024. "Name-Based Estimators of Intergenera-

tional Mobility." *The Economic Journal* 134 (663): 2982–3016.

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.

Appendix A: 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, in table A1, I report 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 third 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 A2 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 A3 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 A4 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 A1: Intergenerational mobility estimates

	Rank rank correlation	IGE	Personal income rank (daughters)	Personal income rank (sons)
	(1)	(2)	(3)	(4)
Coefficient	0.324	0.324	0.291	0.291
	(0.001)	(0.001)	(0.001)	(0.001)
N	1,703,392	1,703,392	866,627	829,039
R^2	0.105	0.092	0.096	0.096

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. Standard errors are in parentheses.

Table A2: Intergenerational mobility estimates: varying years of income

Years of income	1	2	3	4	5	6	7	8	9
Coefficient	0.275	0.288	0.302	0.306	0.313	0.316	0.321	0.323	0.324
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
N	169,594	169,594	169,594	169,594	169,594	169,594	169,594	169,594	169,594
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 A3: Intergenerational mobility estimates: measuring child income at different ages

	(1)	(2)	(3)
Coefficient	0.274	0.304	0.309
	(0.002)	(0.002)	(0.002)
Age child	30-33	34-37	38-41
N	326,420	326,420	326,420
R^2	0.076	0.093	0.096

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 A4: Intergenerational mobility estimates: measuring parent income at different ages

	(1)	(2)	(3)
Coefficient	0.290	0.295	0.295
	(0.001)	(0.001)	(0.001)
Years of income measurement parents	2003-2007	2006-2010	2009-2013
N	1,268,364	1,268,364	1,268,364

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 B: supplementary results

Table B1: Descriptive statistics for the income analysis sample

	Mean	SD	Mean	SD	% missing
Characteristics children					
Year of birth	1984.6	2.9			0.000
Male	0.51	0.50			0.000
Family size	2.70	1.32			0.000
Household income	102,116	65,337			0.000
Second generation migrant	0.15	0.36			0.000
Third generation migrant	0.06	0.23			0.000
Family characteristics: measured at the household level					
Household income rank	50.0	0.29			0.009
Share of primary income parents	0.79	0.27			0.011
Highest education	12.9	3.6			0.358
Total wealth rank	50.0	0.29			0.004
Bank and savings balances	61,287	154,630			0.004
Bonds and shares	34,437	310,763			0.004
Substantial interest	68,213	1,205,399			0.004
House value	328,125	299,765			0.004
Other real estate	36,258	302,694			0.004
Entrepreneurial assets	15,664	129,451			0.004
Other assets	8,231	115,015			0.004
Total debt	159,740	372,346			0.004
Mortgage debt	142,385	181,657			0.004
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	68,151	51,446	29,185	21,736	0.108
Personal earnings	83,105	61,809	33,191	26,961	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
Entrepreneurship	0.116	0.288	0.066	0.218	0.055
Entrepreneurship (other)	0.005	0.051	0.03	0.123	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	0.065	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	0.201	0.356	0.195	0.338	0.066
Years of education	12.8	3.8	11.9	3.7	0.53
Average hourly wage	32.0	26.9	20.7	18.1	0.315
Most important sector of employment (68 categories)	-	-	-	-	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	0.157	0.006	0.079	0.014
Suspected of other crime	0.042	0.2	0.012	0.11	0.014
Total health costs	2,693	7,144	2,559	8,116	0.063
General practitioner costs	174	143	197	155	0.063
Mental health care costs	234	3,540	321	3,947	0.063
Hospital health care costs	1,830	6,722	1,692	5,012	0.063
Pharmaceutical care costs	527	2,230	542	2,083	0.063
Dental care costs	46	303	44	299	0.063
Year of birth	1953.6	5.6	1956	5.0	0.009
Age at first birth	29.3	5.5	27.0	4.4	0.000
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	0.000
Father/mother dead	0.008	0.086	0.004	0.065	0.019
Father/mother presence	0.857	0.35	0.962	0.191	0.037
Extended family outcomes					
Average income rank	0.50	0.22	0.50	0.22	0.246
Average education	12.6	3.2	12.7	3.1	0.420
Average wealth rank	0.51	0.23	0.51	0.23	0.238
Average health expenditures	2,717	5,537	2,565	5,371	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. 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 and do not differentiate by gender.

In addition, I identify the primary sources of personal income, classified into 11 categories. 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. For each parent and each type of asset or liability, I first calculate the average value over the years 2006 and 2011. Subsequently, for each child, I determine the mean of the father's and mother's averaged values for each asset or liability type during this period.

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 (i.e., 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.

Education. Parents' years of education are based on the conversion table in Appendix D1. Table B1 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 3, 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 B2: 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	3,408	0.163	0.141	0.185
0.02	6,817	0.155	0.141	0.171
0.05	17,041	0.158	0.148	0.168
0.1	34,082	0.162	0.154	0.169
0.2	68,163	0.165	0.159	0.170
0.4	136,326	0.167	0.163	0.170
0.6	204,488	0.165	0.162	0.168
0.8	272,651	0.164	0.161	0.166

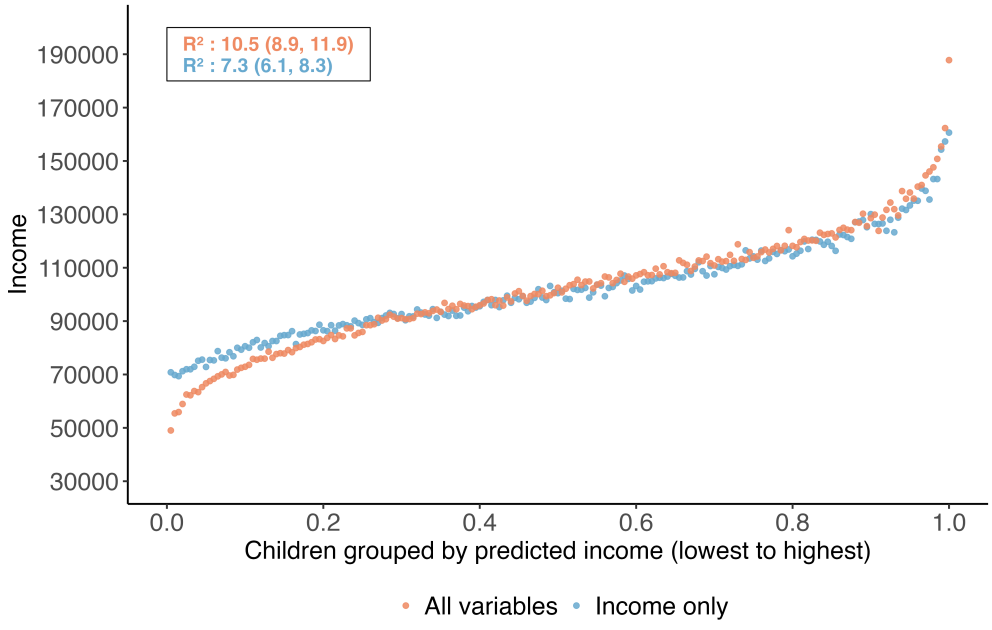
Notes: This table presents estimates of explanatory power for gradient-boosted decision trees that include all explanatory variables (as in Figure 2), using smaller samples. Column 1 reports the share of the core sample that is used for the analysis. Column 2 reports the sample size of the test-data. 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 B3: 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	0.135	0.130	0.140
2	0.141	0.136	0.146
3	0.147	0.142	0.153
4	0.149	0.144	0.154
5	0.167	0.162	0.172
6	0.154	0.148	0.159
7	0.158	0.153	0.163
8	0.158	0.153	0.163
9	0.160	0.155	0.166
Age child	B. Varying ages of income measurement		
30-33	0.125	0.120	0.130
34-37	0.151	0.146	0.156
38-41	0.153	0.148	0.159

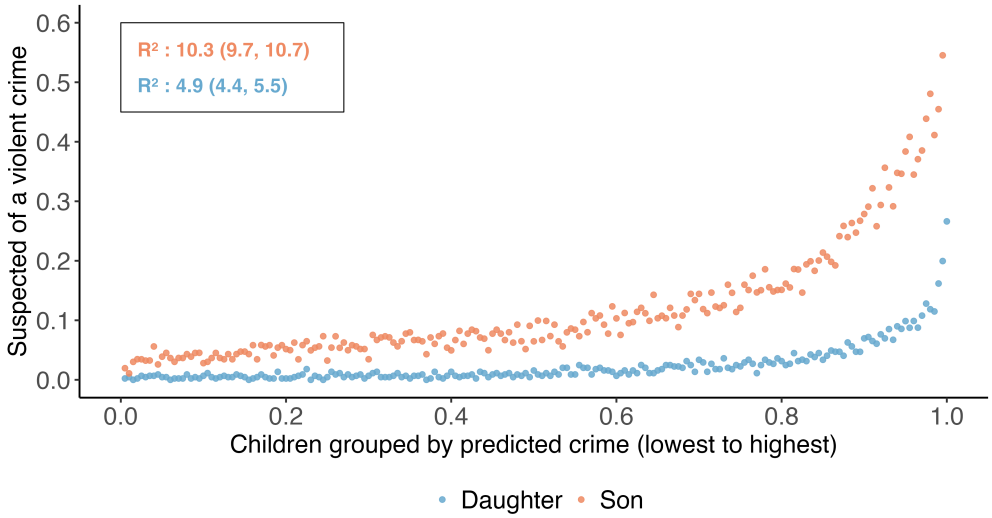
Notes: Each row presents the R^2 and corresponding 95% lower and upper bound for gradient-boosted decision trees that include all explanatory variables to predict child income (as in Section 5). Panel A varies the number of years of income data used to construct the child income rank. Panel B always uses 4 years of income data, but varies the ages at which income is measured. 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%. Confidence intervals for the R^2 are bootstrapped from the test-data using 599 draws.

Figure B1: Predicting children's income level



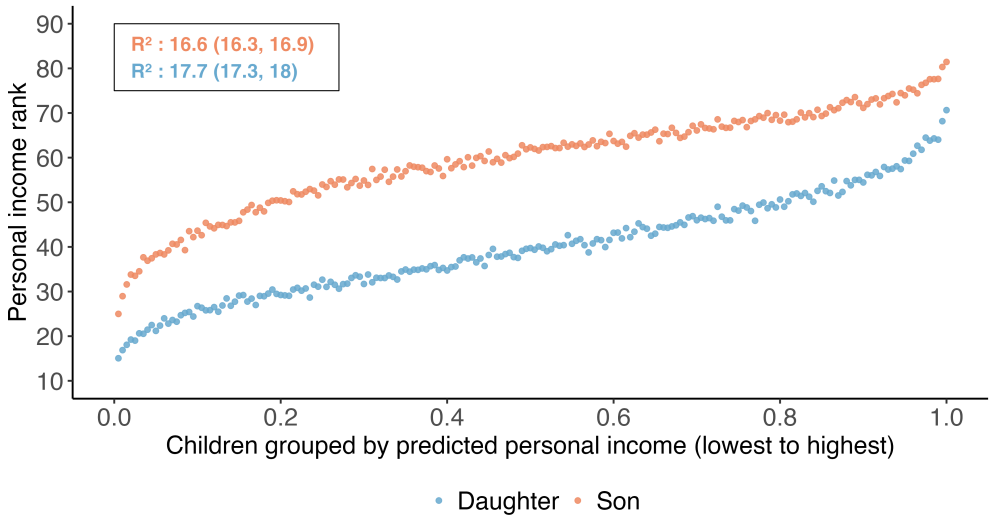
Notes: this figure presents binscatter plots of children's household income, who are sorted into bins based on their predicted income rank. The graphs are constructed using the same sample and steps as in Figure 4, applied to children's income levels instead of ranks. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and are reported in brackets

Figure B2: Predicting children's violent crime by gender



Notes: this figure presents a scatter plot of predicted crime for 92,725 sons and 89,051 daughters separately. Crime is measured as an indicator that equals 1 if a child was suspected of a violent crime between the ages of 20 to 33. The graphs are constructed using the same steps as in Figure 4. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and are reported in brackets.

Figure B3: Predicting children's personal income by gender



Notes: this figure presents binscatter plots of sons' and daughters' personal income ranks for 173,486 sons and 165,769 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 4, 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 B4: Predicting brother-sister differences in income and education

	A. Predicting brother-sister differences in income ranks. R^2 : 0.13 (0.06, 0.19).									
Income gap	-0.9	-0.9	-1.1	-3.3	-2.3	-2.8	-3.3	-3.5	-4.6	-4.8
Income rank brother	49.5	50.1	50	49.6	50.6	50.4	50.1	49.3	48.3	48.5
Income rank sister	50.4	51	51	52.9	52.9	53.2	53.3	52.8	52.9	53.3
Parental income rank	59	57.3	55.8	55.5	54.7	52.5	50.1	48	44.6	41.9
Income rank father	65.6	62.7	60.4	58.5	56.1	52.3	48.2	43.2	36.3	25.5
Income rank mother	40.2	41.1	42	43.9	46.8	49.5	52.5	57.4	59.8	63.1
N	5,621	5,624	5,623	5,625	5,624	5,622	5,624	5,624	5,623	5,630
	A. Predicting brother-sister differences in education. R^2 : 0.28 (0.12, 0.43).									
Education gap	-0.32	-0.32	-0.48	-0.46	-0.62	-0.45	-0.62	-0.61	-0.77	-0.89
Education brother	13.63	14.07	14.25	14.37	14.44	14.73	14.51	14.49	14.41	14.34
Education sister	13.95	14.40	14.73	14.84	15.07	15.19	15.13	15.10	15.18	15.22
Parental income rank	49.7	54.6	56.3	57.8	58.2	58.9	57	54.3	50.6	45.1
Income rank father	54.8	58.4	58.9	58.9	57.8	57	53.1	46.6	41.4	33.6
Income rank mother	33.1	35.2	39	43.4	47.4	53.4	58	63.2	67.5	66.7
N	2,046	2,048	2,046	2,048	2,048	2,048	2,046	2,048	2,046	2,050

Notes: Each column presents descriptive statistics for families grouped into bins based on their predicted income or education gap between sons and daughters. The predictions are generated using the following steps: (i) select all families with at least one son and one daughter, (ii) compute the average income or education difference between them, (iii) randomly sample 80% of families to train the same machine learning model used in the main results to predict these differences based on all family background characteristics. The table reports statistics for the remaining 20% of families, sorted into bins from the lowest to highest predicted income (Panel A) or education (Panel B) difference. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and are reported in brackets.

Table B5: The effect of family background on income: regression results with adoptees

	(1)	(2)	(3)	(4)	(5)
Predicted income	0.275***	0.276***	0.287***	0.290***	0.385***
	(0.026)	(0.025)	(0.026)	(0.026)	(0.072)
Parental income rank					-0.050
					(0.047)
Controls		x	x	x	
Country of Origin FE			x	x	
Year of Adoption FE				x	
N	4,938	4,938	4,938	4,938	4,938
R^2	0.008	0.024	0.044	0.045	0.046

Notes: Each column shows results from a separate regression of a child's income rank on predicted income. Controls are a gender dummy and age-at-migration. The predicted values for income are based on gradient-boosted decision trees reported in Figure 2. The fixed effects are fully interacted. Standard errors are in parentheses. (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

Table B6: Descriptive statistics for international adoptees and their parents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Predicted income rank	37.9	45.4	49.3	51.9	54.4	56.6	58.7	61.1	63.8	69.3
Income rank	36.6	36.0	39.9	38.2	41.8	41.8	43.4	40.2	44.4	43.2
	Characteristics Adoptive Parents									
Parental income rank	20.2	30.1	36.5	44.4	52.5	60.2	68.1	77.3	84.4	93.1
Parental wealth rank	29.3	43.9	53.7	58.1	63.3	67.9	70.3	70.0	74.0	82.3
Highest education parents	11.2	12.0	12.8	13.6	14.2	14.3	14.9	15.4	16.1	16.5
Father suspected of crime	0.08	0.05	0.04	0.04	0.05	0.03	0.02	0.02	0.02	0.02
Health expenditures parents	4,554	3,993	3,220	2,754	3,290	2,570	2,851	2,742	2,163	2,569
Extended family income rank	38.5	44.8	47.6	49.0	53.1	56.3	58.5	59.0	63.4	70.3
N	493	494	494	494	494	493	494	494	494	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 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 $f(x_1, ..., x_n)$ is generated from a set of variables $x_j \in X$. 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.

Lundberg and Lee (2017) show that Shapley values are the only measures of variable importance that preserve important properties from cooperative game theory.[44] 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 polynomial time, making these explanations feasible (Lundberg et al. (2020)).

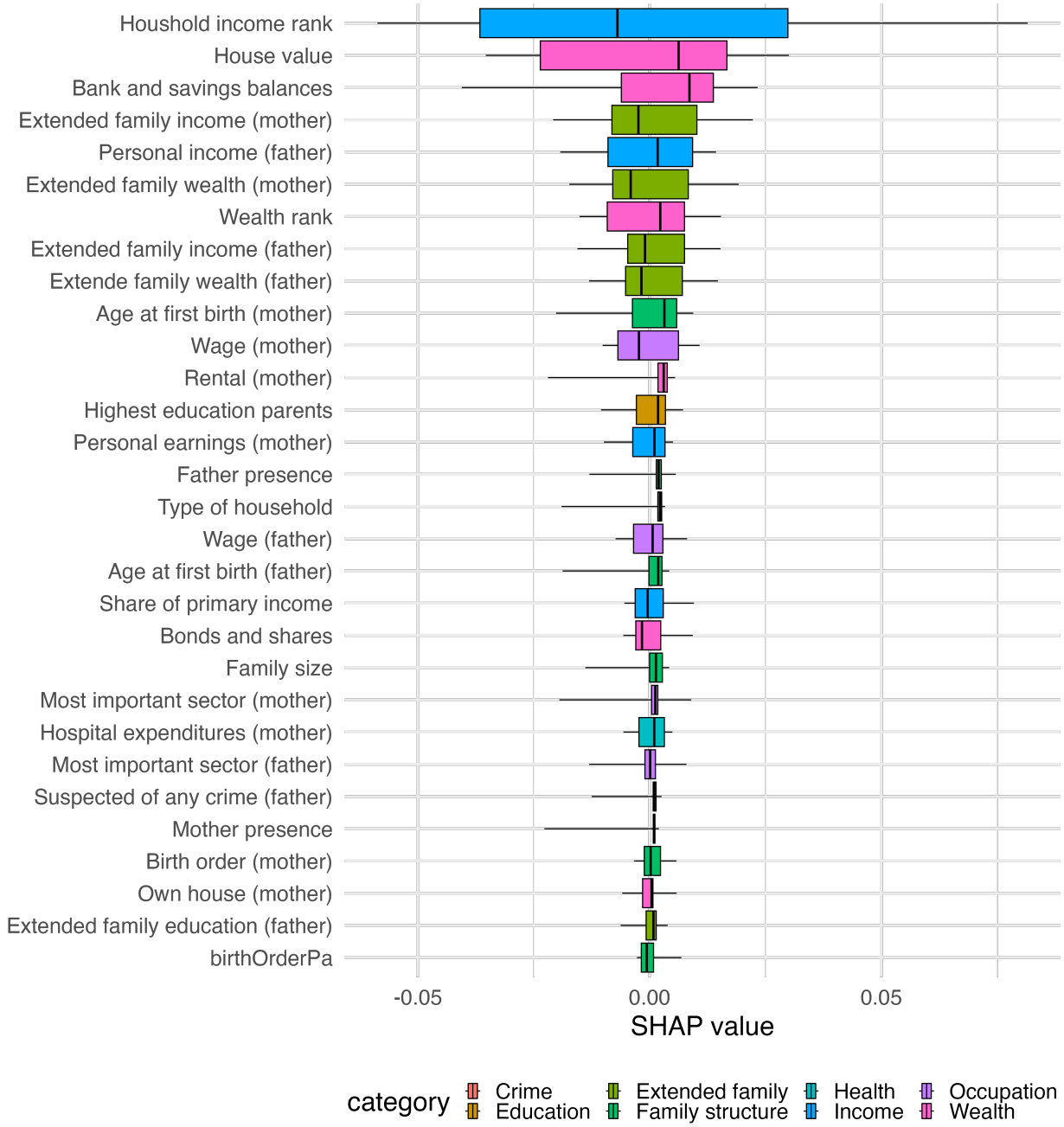
Using this algorithm, I compute Shapley values for the gradient-boosted decision tree model used in the main results (2), applied to a random sample of 10,000 children from the test dataset. This process generates Shapley values for each variable and each child. Figure C1 presents a boxplot of Shapley 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 -0.06, indicating that for 2.5 percent of the children, parental income reduces the prediction by at least 0.06 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. Generally, of all explanatory variables, only the indicators for whether the father or mother is identified in the data provide no contribution to the predictions.

[44] 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 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 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 D1 provides a simplified overview of the levels of education and their corresponding years of schooling. The abbreviations are explained in Table D1. 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 D1 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.

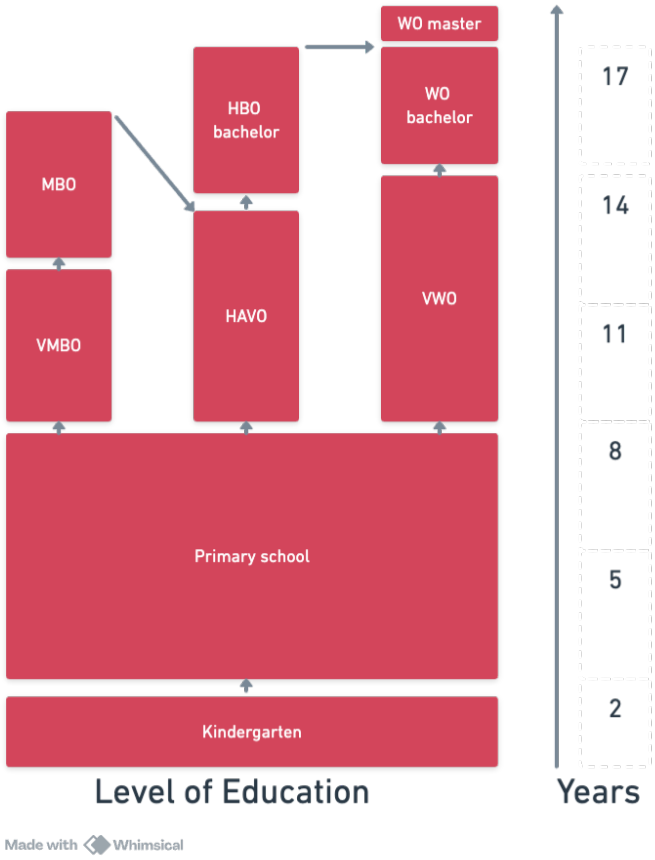


Figure D1: The Dutch Educational System

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 D1: 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