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

Sander de Vries*

July 22, 2025

Abstract

Understanding the importance of family background for children's economic success is essential for evaluating inequality and informing policy. This paper provides new insights by linking key economic outcomes of 1.7 million Dutch children to exceptionally rich information about their parents and aunts and uncles. I provide a detailed analysis of how these outcomes vary across family backgrounds and show that incorporating all information explains substantially more variation in income (+58%), education (+107%), and violent crime (+169%) than analyses using parental income only. I also present new evidence on the role of neighborhoods and pre-birth factors in driving income inequalities.

Keywords: intergenerational mobility, inequality of opportunity

JEL Codes: I24, J24, J62

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

1 Introduction

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

Yet, the most commonly used measures of intergenerational mobility provide limited insights on how income is distributed across family backgrounds. Estimates based on parent-child income associations capture just one dimension of family background and may understate its broader influence. To address this limitation, researchers also frequently estimate sibling correlations (Solon (1999)). However, these correlations capture all shared influences between siblings, including factors unrelated to parents, and may greatly overstate the role of the family (Collado et al. (2023)). Consequently, the extent to which family background contributes to income inequality remains unclear. Moreover, neither approach sheds light on which specific characteristics beyond parental income distinguish families whose children experience high or low income prospects.

The main contribution of this paper is to provide an exceptionally detailed analysis of how children's incomes are distributed across family backgrounds. I link the long-run incomes of 1.7 million Dutch children to detailed information on their fathers, mothers, and aunts and uncles, including income, wealth, occupation, education, criminal records, health, migration background, and family structure. Although this dataset does not capture all family background dimensions, it enables me to study disparities between families in much greater detail than prior work. I use all family information in a flexible machine learning model to predict child income ranks. This predictive model provides a non-parametric description of how children's expected income ranks vary by their family background characteristics.

Incorporating all family information reveals substantially greater disparities between families than analyses based on parental income alone. While parental income explains 10.5 percent of the variation in child income ranks, the full set of family characteristics accounts for 16.6 percent, a 58 percent increase in explanatory power. To illustrate, the 0.5 percent of children with the lowest expected incomes based on parental income alone have an average observed income rank of 31; under the comprehensive model, this drops to 18. To the best of my knowledge, no other paper has identified individuals with such low expected income ranks based solely on family background. The increase in explanatory power is even larger for predicting children's education (107%) and crime (169%). These findings highlight the value of a multidimensional approach for accurately quantifying family-driven inequalities.

Prior papers often focus on aggregated statistics, such as intergenerational mobility coefficients or inequality of opportunity estimates. A second contribution of this paper is to move beyond these aggregate measures by providing granular insights into the distribution of children's expected incomes and the corresponding family characteristics. For instance, I show that the 0.5 percent of children with the highest expected incomes have an observed mean income rank of 78 and come from families with favorable characteristics across nearly all dimensions, with particularly high parental income and wealth. In contrast, those from the bottom 0.5 percent (mean rank 18) have parents who are often young, separated, and who have low income and wealth, limited education, poor health, and criminal records, with similar disadvantages among aunts and uncles. Overall, parental and extended family income and wealth emerge as the best predictors of children's income, underscoring the central role of income and wealth data in quantifying intergenerational dependence.

Lastly, I examine the roles of neighborhoods and pre-birth factors. Neighborhoods contribute little to income inequality: while there is considerable sorting into neighborhoods, neighborhood differences account for only a small share of income disparities across family backgrounds and they explain little income variation beyond what is already explained by the observed family characteristics. Instead, much of family (dis)advantage appears to

originate before birth. Evidence from international adoptees shows that being raised from infancy in an advantaged family increases income, but considerably less than for own-birth children.

This paper contributes to multiple strands of literature studying intergenerational mobility or the importance of family background more generally. 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 (Black and Devereux (2011)). 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.

Only a small number of studies also relate children's long-run incomes to multiple family background characteristics. Most closely related are recent papers using machine learning to predict children's income (Blundell and Risa (2019), Brunori et al. (2023), Brunori et al. (2024), Chang et al. (2025)).[1] The latter three papers use much smaller survey datasets and include a limited number of parental background characteristics. Blundell and Risa (2019) use administrative data with information about parents' income, net taxable wealth, education, occupation, marital status, family size and neighborhoods. This paper uses far more detailed data, including the value of specific types of assets and debt, health expenditures, criminal behavior, family structure, migration background, and extended family outcomes, all of which are shown to be significant predictors.[2] Moreover, Blundell and Risa only report explanatory power estimates. This paper provides a substantially more granular analysis

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

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

by reporting the full distribution of expected income alongside the corresponding family background characteristics.

Sibling correlations are often viewed as an appealing alternative to measuring the importance of family background (Solon (1999)). However, this paper highlights two advantages of the prediction approach compared to sibling correlations. First, sibling correlations also capture unobserved factors unrelated to parents, such as sibling spillovers or shared external shocks, complicating interpretation.[3] The prediction approach instead relies exclusively on observable family characteristics, whose contributions can be quantified.[4] Second, sibling correlations measure only overall explanatory power, whereas the prediction method generates a complete distribution of expected incomes. This distinction is important, as equity implications may vary depending on where along the income distribution family background has the greatest influence (Roemer and Trannoy (2016)).

Lastly, this paper presents new evidence on intergenerational mobility in the Netherlands. While recent studies report rank-rank correlations between 0.16 and 0.23 (Van Elk et al. (2024), Manduca et al. (2024), Boustan et al. (2025)), in my baseline analysis, I obtain a substantially higher estimate at 0.32. This places the Netherlands among the Western countries with relatively high intergenerational income persistence. I discuss in detail why our estimates differ.

This paper proceeds as follows. Section 2 presents a simple theoretical framework, linking the approach in this paper to other commonly used methods for quantifying intergenerational dependence. Section 3 presents the data. Section 4 discusses model estimation and inference. Section 5 provides the main results. Section 6 explores mechanisms. Section 7 concludes.

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

[4]In section 5.5, I show that the observable family background characteristics explain about half of siblings' income similarities. The remaining half is driven by unobserved factors uncorrelated with these observables.

2 Models of Intergenerational Dependence

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

Let Y_{sf} be the income of a child s in a family f and let Y_f be parental income. Let $\mathbf{X}_f = (Y_f, X_{f1}, ..., X_{fk}) \subset \mathcal{X}$ be the observable family background characteristics that siblings share and $\mathbf{Z}_f = (Z_{f1}, ..., Z_{fl}) \in \mathcal{Z}$ be the unobservables factors that siblings share.[5] Consider the following two conditional expectation function decompositions of Y_{sf}:[6]

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

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 the observable and unobservable factors. In the observables model, the groups include all children sharing the same observable family characteristics.

The primary objective of this paper is to measure the importance of observable family background characteristics for children's income. I quantify this by the share of income variation attributable to differences in $g(\mathbf{X}_f)$ — the conditional mean for individuals with

[5] 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.

[6] 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.

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'.

Although the sibling model f depends on unobservables, its explanatory power is identified when sibling links are observed in the data. This is because the correlation between two randomly drawn siblings coincides with the model's explanatory power (Solon (1999)):

$$\begin{aligned} Cor(Y_{sf}, Y_{cf}) &= \frac{Cov(Y_{sf}, Y_{cf})}{V(Y_{sf})} = \frac{Cov(f(\mathbf{X}_f, \mathbf{Z}_f) + e_{sf}, f(\mathbf{X}_f, \mathbf{Z}_f) + e_{cf})}{V(Y_{sf})} \\ &= \frac{V(f(\mathbf{X}_f, \mathbf{Z}_f))}{V(Y_{sf})} = R^2_{y|f} \end{aligned} \tag{4}$$

This expression shows that the sibling correlation captures all factors siblings share, both observed and unobserved. It therefore also provides an upper bound on the explanatory power of any predictive model using only observed characteristics shared between siblings.

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 both f and g will have little explanatory power. Still, however, the observables model g can be a good approximation of the sibling model f. In that case, the fraction of the variance in $f(\mathbf{X}_f, \mathbf{Z}_f)$ that is explained by $g(\mathbf{X}_f)$,

$$R^2_{f|g} = \frac{V(g(\mathbf{X}_f))}{V(f(\mathbf{X}_f, \mathbf{Z}_f))} = \frac{R^2_{y|g}}{R^2_{y|f}}, \tag{5}$$

is high. I call this fraction the *sibling completeness*.

A sibling completeness close to zero means that siblings' similarities arise from factors

uncorrelated with the observables. On the other hand, if the sibling completeness is close to one, then the observable family background characteristics are nearly as predictive as the model that also includes unobservable factors shared between siblings.[7] To illustrate these ideas, I report estimates of the sibling correlation and the sibling completeness of a comprehensive model predicting child income ranks in Section 5.5.

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.[8] 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.

[7] 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.

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

3 Data

Core analysis sample. I use administrative data from Statistics Netherlands covering the full Dutch population.[9] The main sample consists of all children born in the Netherlands between 1980 and 1989, excluding 3.4 percent with missing income data, resulting in 1,703,038 observations.

The main outcome in this paper is a child's long-run gross household income rank. I focus on household income because it provides a reliable measure of economic resources even in the case of non-participation in the labor market and it is commonly used in other intergenerational mobility studies (Chadwick and Solon (2002)). Nevertheless, I also present results using personal income ranks to abstract away from household formation considerations. Household incomes are observed between 2003 and 2023 and includes income from employment, entrepreneurship, capital, income insurance payments, social security payments, inter-household income transfers (such as alimony), and contributions to social insurance made by both employers and employees.[10] I measure income in 2024 euros, adjusting for inflation using the consumer price index.

I construct a proxy for children's lifetime household income by averaging their household income from age 30 onward.[11] This approach reduces measurement error from transitory income shocks (Mazumder, 2005) and life-cycle bias (Haider and Solon (2006), Nybom and Stuhler (2017)). I observe income up to age 43 for the oldest cohort (born in 1980) and up to age 34 for the youngest cohort (born in 1989). On average, children have nine income observations, with 96 percent having at least five. I then rank children within birth-years based on their lifetime household income. I also present results for various alternative measures to evaluate the sensitivity of the results due to these choices.

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

[10] 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.

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

Children's education and crime. The education register contains individuals highest attained education. I use this register to construct a years-of-education variable according to the conversion table in Appendix D. Statistics Netherlands began comprehensive recording of education across all levels only in the early 2000s, resulting in significantly reduced coverage for cohorts born before 1985. As such, I limit the analysis to the subsample of children born from 1985 to 1989. After excluding 0.5 percent with missing education records, the education sample includes 908,876 children.

The crime register data contains all offenses reported to the police between 2005 and 2022, including the individual identifier of the suspected offender(s). The crime outcome is an indicator of whether a child has been suspected of any *violent* crime at ages 20 to 33. I focus on violent crime because of its high societal costs. This age range represents the longest window for which I can accurately observe children's criminal behavior and coincides with prime years of criminal activity. Because crime is high in the early twenties, this analysis is also restricted to children born between 1985 and 1989, for whom complete crime histories are available over this age span. The analysis focuses on boys only, resulting in 463,625 observations.

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

Other explanatory variables. Table 1 describes how the other variables are classified into eight 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 8 categorical variables (two containing 68 distinct categories and six containing 8 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.

Table 1: Explanatory Variables

Income	Household income, personal income, personal earnings, most important sources of personal income (in 11 categories), and the primary household income share.
Wealth	The value of bank and savings balances, bonds and shares, real estate, entrepreneurial assets and liabilities, other assets, mortgage debt, study debt, and other debt.
Occupation	Average hourly wage and most important sector of employment (in 68 categories).
Education	Highest level of completed education.
Health	Average healthcare costs for 5 categories*: general practitioner, hospital, pharmaceutical, mental health care, and dental care.
Crime	Indicators of whether the parent has been suspected of a property, violent, or other type of crime.
Family structure	Parents' family size, age-at-first-birth, birth order, single-parent household, father or mother presence, parental death, child family size, and whether the father or the mother are identified.
Migration background	Region of origin of the father, mother, and all grandparents (in 8 categories).
Extended family outcomes	Average years of education, household income rank, wealth rank, total healthcare costs, and share of all siblings of the parent who have been suspected of a crime.

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

Although the data are rich, they come with two limitations. First, some parental outcomes are observed only after their children have left the household. Consequently, my results may underestimate the importance of family background compared to a model that

includes information on parents' resources and well-being during their children's formative years. Nonetheless, many parental characteristics are highly persistent over the life cycle, making them a reasonable proxy for the family environment at earlier ages.[12]

Second, despite the extensive coverage of variables, some missing values persist. Most importantly, education records for the parents' generation are incomplete. In the robustness check, I assess the impact of these missing education records. 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.

4 Model Training and Evaluation

The objective is to train a predictive model that accurately predicts the conditional expectation function $g(\mathbf{X}_f)$ of children's income given the family background characteristics (see Equation 2). A key challenge is that the true functional form of $g(\mathbf{X}_f)$ is unknown. Variables may enter in a non-linear manner or interact with other variables. In these cases, flexible machine learning methods can outperform linear regression models. Accordingly, I employ gradient-boosted decision trees to generate these predictions (Friedman (2001)). Tree-based methods offer the additional advantage of providing Shapley value–based measures of variable importance even with a large number of predictors, which is infeasible with most alternative methods (Lundberg and Lee (2017), Lundberg et al. (2020)).

For each analysis requiring a separate predictive model, I randomly split the sample into a training set (80 percent) and a test set (20 percent). I perform 5-fold cross-validation on the training data to determine the optimal parameter values, and then train a final model on the full training set using these parameters. This model is then applied to observations from

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

the test data to estimate the out-of-sample explanatory power (R^2). Generally, all results in this paper that rely on predictions are based on observations from the test data.

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.

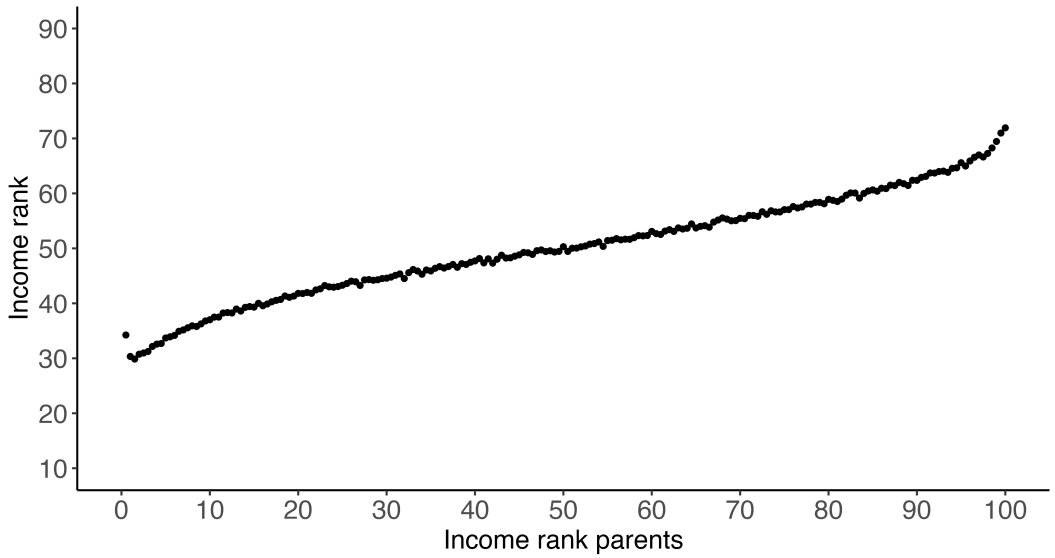
Figure 1 presents a scatter plot of children's income ranks relative to their parents' income ranks. The X-axis is divided into 200 bins, each representing half a percentile and containing roughly 8,500 children. The dots correspond to the mean household income rank of children given their parents' household 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.[13] 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.

This estimate positions the Netherlands among the OECD countries with relatively strong persistence. Intergenerational persistence in the Netherlands is higher than in Sweden, Denmark, Australia, Norway, Germany, and Canada (0.18-0.24), similar to France, Italy, and the UK ($\sim$ 0.30), and lower than in the US (0.36).[14] Despite the Netherlands' relatively low-income inequality and accessible education, intergenerational persistence appears surprisingly high. This should be taken into account when interpreting the results of subsequent analyses that incorporate additional family background information. These results may be

[13] As noted before by Van Elk et al. (2024), there is some measurement error at the very bottom of the parental income distribution. This is because some wealthy parents report low income as a result of capital losses. Removing the bottom 0.5 percent of the sample does not affect the estimates much.

[14] 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). See Kenedi and Sirugue (2023) for a more detailed comparison of approaches.

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 $N = 1,702,355$ children. The X-axis reports the parent income rank sorted into 200 equal-sized bins. The Y-axis reports the mean income rank within each bin.

more likely to generalize to countries with similarly high intergenerational persistence as the Netherlands.

The estimated rank-rank correlation exceeds recent estimates from Van Elk et al. (2024), Manduca et al. (2024), and Boustan et al. (2025), who report estimates between 0.16 and 0.23. As estimating the rank-rank correlation is not the main focus of this paper, I defer a detailed discussion of why our estimates differ to Appendix B.

Appendix B also reports additional, commonly used mobility estimates to facilitate cross-country comparisons. Moreover, I vary the number of years over which parental income is measured and the timing of income measurement in parents' and children's lives. These robustness checks suggest that the estimate is robust to measurement error and lifecycle bias. This further implies that the explanatory power of additional variables is unlikely to reflect mere corrections for measurement error in parental lifetime income.

5.2 Including Detailed Parental Information

The previous section analyzed how children's incomes vary with parental income. A key advantage of this approach is its comparability, as similar estimates are available for many other countries. However, as noted in the introduction, parental income is only one dimension of family background, leaving unclear the extent to which the broader family background shapes income inequality. To shed light on this, this section examines how children's incomes vary across a much broader set of family characteristics.

To quantify the increase in family-driven inequality when adding the broader family background information, I compare the explanatory power of a model using only parental income with that of a model incorporating all explanatory variables. Both models are trained and evaluated on the same training and test data. For the income-only model, I non-parametrically predict a child's income rank in the test data by the mean income rank of all children in the training data with the same parental income rank and year of birth. Like the linear regression in the previous section, this model achieves an explanatory power of 10.5 percent. The predictions using all explanatory variables are generated by a tuned gradient-boosted decision tree, as described in Section 4. This model includes all explanatory variables from Table 1 and children's year of birth.

Adding all information about the parents reveals substantially stronger intergenerational dependence. The comprehensive model achieves an explanatory power of 16.6 percent, marking a 58 percent increase compared to the income-only model. To put this into perspective, an increase in the rank-rank correlation from 0.32 to 0.41 would result in the same increase in R^2.[15] While this may seem modest, it is considerable, considering the difference in rank-rank correlation between Denmark (high mobility) and the US (low mobility) is about 0.16 (Helsø (2021), Davis and Mazumder (2024)). Moreover, the increase in R^2 far exceeds the gain achieved from reducing attenuation bias in an income rank-rank regression.[16] This

[15] I use here that in a rank-rank regression, $R^2 = \beta^2$ (i.e. $0.408^2 - 0.324^2 = 0.166 - 0.105 = 0.061$).

[16] Table B2 columns 1 and 9 shows that using 9 years of income data versus one year of income data in a rank-rank regression increases the R^2 from 8.2% to 10.0%.

source of measurement error has received considerable attention in the literature (Mazumder (2005), Nybom and Stuhler (2017)). When the goal is to quantify income disparities between families, adding more information about parents is thus more valuable than constructing a more accurate proxy of lifetime income.

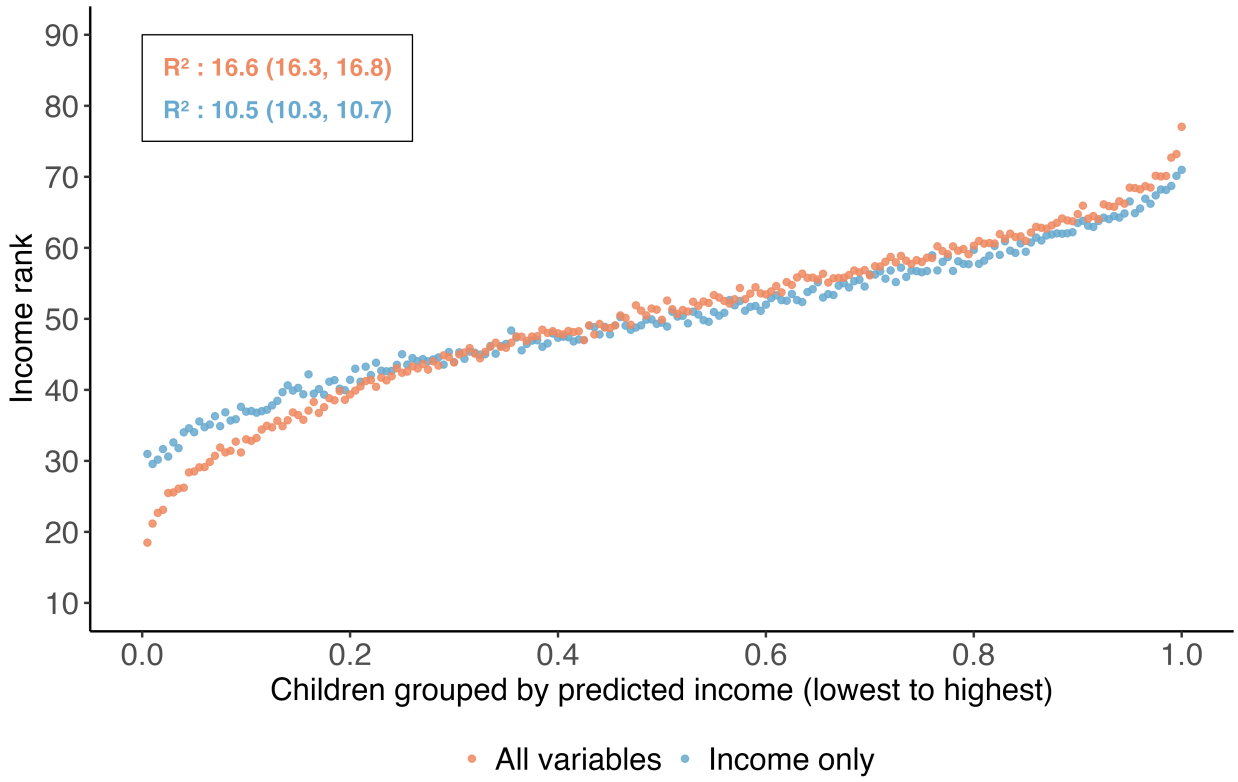
Figure 2 provides a binscatter plot of children's mean income ranks, sorted from lowest to highest predicted income. The X-axis divides the test dataset into 200 bins, each containing approximately 1,700 children, based on their predicted income ranks within their cohort. The Y-axis reports the average observed income rank for each bin. The blue dots represent children grouped by predicted income using parental income alone, while the orange dots reflect groupings based on predictions from the comprehensive model.

The comprehensive model identifies considerably greater income disparities by family background. For instance, in the income-only model, the 0.5 percent of children with the lowest expected income have an average income rank of 31. With the comprehensive model, this drops to 18. Similarly, for the top 0.5 percent, the income-only model estimates an average rank of 70, while incorporating additional family background information raises this to 78. To the best of my knowledge, no other study has identified children with similarly low or high expected income ranks based solely on family background information.

The results also reveal large differences within the top and the bottom of the expected income distribution. For example, the average income of those in the top 0.5 percent is 4 ranks higher than that of the next 0.5 percent. Similarly, the average income of the bottom 0.5 percent is 3 ranks lower than that of the next 0.5 percent. These differences are considerably larger than when using parental income, in which case the largest income difference between any two consecutive bins is 1.5 ranks.

Overall, family background has a stronger influence on child income than analyses based solely on parental income suggest. Since support for redistributive or opportunity-equalizing policies depends on the perceived role of family background in shaping economic outcomes (Alesina et al. (2018)), these findings provide important evidence for the public debate.

Figure 2: Predicting Child Income with Detailed Parental Information



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

5.3 What Characterizes Family (Dis)Advantage?

The previous section documents substantial income disparities across families. This section examines the key predictors of these disparities and describes the characteristics of families with the highest and lowest expected child incomes. This provides insight into the sources of strong or weak income prospects and can help improve the targeting of disadvantaged children for policy support.

Table 2 shows how family characteristics vary across the expected income distribution, focusing on the most advantaged and disadvantaged children.[17] The first and last four columns include the 10 percent of children with the lowest and highest expected incomes, while the fifth column contains all children in between. Row 1 shows the corresponding mean income ranks and the remaining rows report the average family background characteristics.

The first four columns of Table 2 show that children at the bottom of the expected income distribution face cumulative disadvantages. They have parents with low income and wealth and who are often young, separated, minimally educated, have high health expenditures, and are often suspected of crimes. Their aunts and uncles also have low income and wealth. These results highlight that policymakers seeking to target the most disadvantaged children should consider more information than parental income only.

Information on multiple family background characteristics can be difficult to obtain. This raises the question: which family characteristics are most important for explaining income disparities? Appendix C presents a detailed graph illustrating the variable importance of the 30 most predictive variables, calculated using Shapley values. This analysis reveals two insights. First, all variables except for whether the father or mother is identified contribute to the predictions, indicating that each adds valuable information to the analysis. Second, income and wealth measures for both parents and extended family are the most influential, with nine of the top ten predictors falling into these categories. This underscores the impor-

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

tance of (i) income and wealth data and (ii) extended family ties in capturing family-driven income inequality.

Table 2: Family Background Characteristics across the Predicted Income Distribution

	Predicted Income Bins								
	0-0.5	0.5-1	1-5	5-10	10-90	90-95	95-99	99-99.5	99.5-100
Child income rank	18.28	21.28	25.78	31.04	50.53	65.82	69.58	73.29	77.58
	Family background characteristics								
Parental income rank	6.32	8.33	11.88	16.28	49.22	87.51	93.24	97.04	98.44
Parental wealth rank	12.11	12.55	14.16	17.32	50.86	74.13	80.33	86.65	89.65
Max. education parents	8.16	8.78	9.52	9.86	13.08	16.10	16.69	17.29	17.44
Health costs parents	5,402	5,136	4,185	3,909	2,596	1,886	1,807	1,711	1,544
Crime father	0.59	0.46	0.32	0.18	0.05	0.02	0.02	0.03	0.03
Extended family income	17.01	20.91	25.46	30.38	49.10	64.5	69.4	74.73	79.46
Extended family wealth	21.96	23.95	26.8	30.92	51.05	63.71	67.82	70.99	74.09
Father presence	0.36	0.34	0.44	0.62	0.88	0.97	0.98	0.98	0.98
Migration background	0.30	0.38	0.49	0.52	0.18	0.10	0.12	0.14	0.15
Age at first birth mother	21.80	22.64	24.11	25.31	27.05	28.39	28.64	28.90	29.00
N	1,703	1,703	13,624	17,030	272,487	17,030	13,624	1,703	1,704

Notes: Each column shows descriptive statistics for a group of children in the test data from the same predicted income bin. 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. All values are averages, with missing values excluded from the calculations. 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. Migration background is an indicator which equals 1 if the child is a second or third generation migrant. The other variables are discussed in Table 1.

5.4 Predicting Education and Crime

I next present results for children's educational attainment and violent criminal behavior. These outcomes are interesting for two reasons. First, education generates substantial private and social returns, while violent crime imposes large societal costs. Understanding the extent to which they are shaped by the family environment is therefore important in its own right. Additionally, considering alternative outcomes allows me to assess whether the value of family background information beyond parental income varies across outcomes, thereby informing the generalizability of the main result.

As violent crimes are predominantly committed by men (85 percent), I focus on men's

criminal behavior only. In Figure A1, I report results for women's criminal behavior too. Moreover, as discussed in Section 3, both analyses are restricted to children born between 1985 and 1989.

Figure 3 (a) reveals strong differences in children's education by family background. For example, children with the 5 percent lowest predicted education levels from the comprehensive model have on average 11.3 years of education, frequently dropping out without qualifications. To the contrary, children with the 5 percent highest predicted education levels have on average 17.1 years of education, corresponding to an undergraduate degree. The explanatory power of this comprehensive model is also markedly higher (+13.5 p.p., 107 percent) than that of the income-only model.[18] This increase in explanatory power considerably exceeds the 60 percent improvement observed when predicting child income (Figure 2).

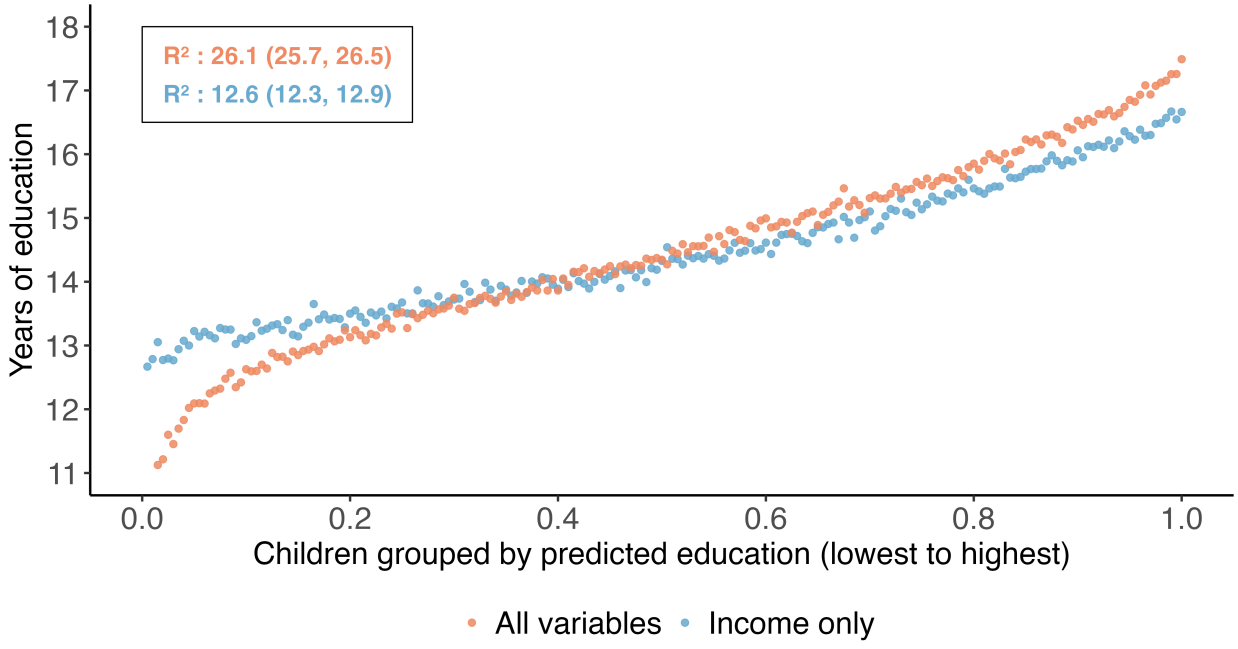
This is also true for predicting boys' violent crime. Figure 3(b) shows that the explanatory power for predicting crime increase from 3.9 percent for the model that incorporates income only to 10.5 percent for the comprehensive model, marking a 169 percent increase. Taken together, it appears that including family dimensions beyond parental income is even more valuable for quantifying disparities in other child outcomes than children's incomes.

Moreover, the results indicate that violent crime is highly concentrated in disadvantaged families. For a small subset of children, the probability of being suspected even exceeds fifty percent. A simple calculation shows that the 20 percent of boys with the highest crime risk in Figure 3 (b) account for 50 percent of all boys who have been suspected of a violent crime between the ages of 20 and 33. These results shed new light on the importance of family background as an important determinant of criminal behavior.[19]

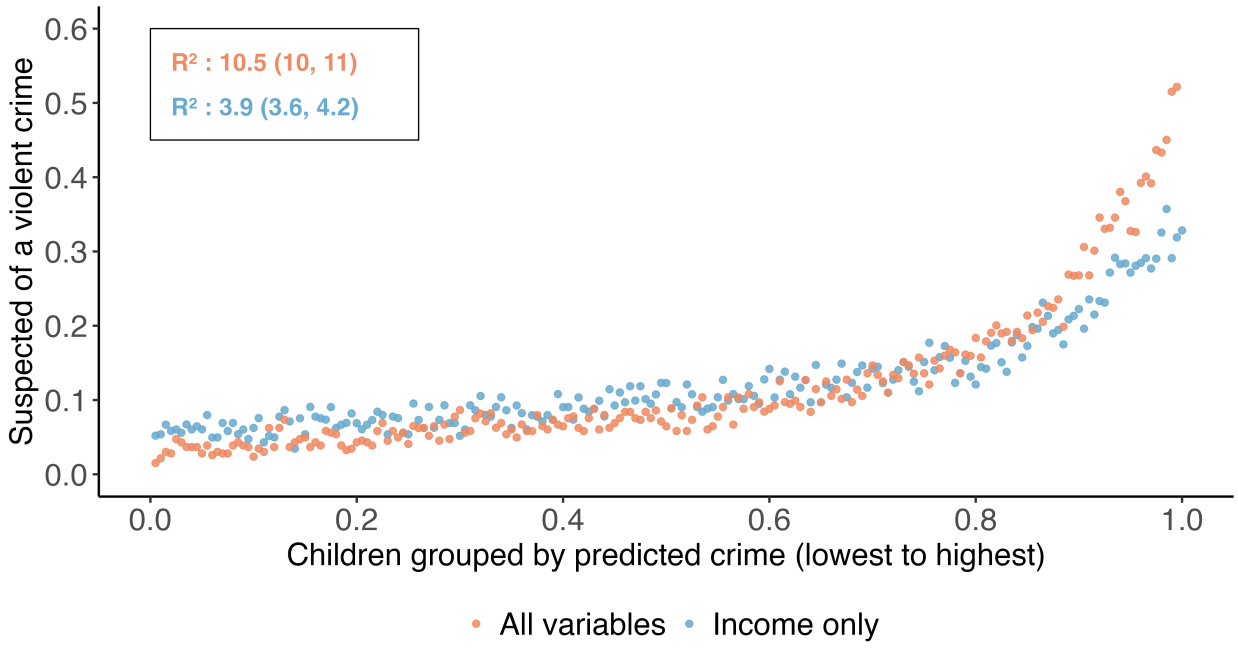
[18]Intergenerational mobility studies often apply regressions of child education on the highest education of the parents. Applying this regression to a subsample of children for whom at least one parent's education is observed, I find an explanatory power of 11.7 percent.

[19]Other works have mostly focused on intergenerational crime associations (see Besemer et al. (2017) for a review) or sibling correlations Eriksson et al. (2016).

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.

5.5 Additional Results

Sibling correlation. A commonly used alternative method to quantify the importance of family background is the sibling correlation, which captures the contribution of all factors shared between siblings (Solon (1999)). I estimate a sibling correlation in income of 0.31 (Table B1). As discussed in Section 2, this provides an upper bound on the explanatory power of any predictive model that solely includes variables that are equal between siblings, as in this paper. The explanatory power of the comprehensive model is about half of this correlation (0.16/0.31), implying a sibling completeness of fifty percent (Equation 5). 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.

Gender differences. Figure A2 presents results from predictive models trained to predict sons' and daughters' household income ranks separately. I also present results using personal income ranks in Figure A3 to abstract away from household formation considerations. The explanatory power for predicting household income ranks is similar between genders, and for predicting personal income ranks, it is somewhat higher for daughters.

Income level differences. The findings above relate to inequalities in gross household income ranks. In Figure A4, I provide results using gross household income and disposable household income in absolute terms, while maintaining the same explanatory variables.[20] The comprehensive model's explanatory power for household income is 11.2%, which is higher than the 8.5% for disposable income. This shows that income redistribution through taxes and transfers considerably diminishes the impact of family (dis)advantages.

Functional form. A straightforward OLS model, which includes all variables linearly,

[20]Disposable income is the amount left after deducting taxes and social insurance payments from gross income.

achieves an explanatory power of 15.3 percent.[21] 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 A2 shows that explanatory power declines with smaller samples but stabilizes once at least 40 percent of the data are used. This suggests that downward bias due to insufficient sample size for training the machine learning model is unlikely.

Table A3 varies the number of years and ages at which child income is measured. Explanatory power attenuates when fewer years of income are used, but stabilizes once about five years of income are used.[22] It also decreases somewhat when income is measured exclusively in the early 30s, but when I re-estimate the model using only incomes beyond age 32, then the overall estimate is virtually identical to the main results specification. This indicates that the influence of attenuation or life-cycle bias is likely minimal.

Finally, I assess the importance of the missing education records. I first train the model on the subset of children whose parents' education is observed ($n = 1,093,245$, $R^2 = 17.4$). I then re-train the model on the same sample after removing all education variables for both parents and extended family ($R^2 = 17.3$). The resulting drop in R^2 is only 0.1 percentage point, indicating that the remaining variables already capture most of the educational variation across families. This suggests that the explanatory power of the model would increase only marginally if complete education data were available.

6 Drivers of Family (Dis)Advantage

Having documented sizable income differences across family backgrounds, an important next step is to ask what explains these gaps. Understanding the underlying mechanisms is essential

[21] Coefficient estimates are available upon request.

[22] In intergenerational mobility regressions, classical measurement error in child income does not bias the coefficient estimate. It only inflates the standard error. However, when estimating explanatory power, such left-hand side measurement error does matter. Reducing measurement error lowers the variance of child income, which in turn affects the explanatory power of the regression.

for informing policies that aim to improve outcomes for children from disadvantaged families.

In this final section, I explore two broad potential channels. First, I explore the role of neighborhoods in driving disparities between families. Second, I present evidence on the importance of post-birth factors in the intergenerational transmission process.

6.1 The Role of Neighborhoods

Wealthier parents often live in more affluent neighborhoods, which may foster child development through better community resources or peer environments. A natural extension of the analysis in Section 5 would be to incorporate neighborhood indicators directly into the predictive model. However, with over 11,000 neighborhoods, this is computationally infeasible. I therefore study the role of neighborhoods separately in this section.

First, I examine the extent to which families systematically sort into different neighborhoods. I regress children's predicted incomes on neighborhood fixed effects, corresponding to the neighborhoods where children are registered at age 15.[23] As the predicted incomes are a measure of family (dis)advantage, a higher explanatory power of this regression means that there is stronger sorting of (dis)advantaged families across neighborhoods. Table 3 column 1 shows that these fixed effects explain 30 percent of the variation in predicted income. This implies that neighborhoods that are one standard deviation more advantaged in terms of family background have average predicted incomes that are 6.1 ranks higher.[24] This indicates substantial sorting of advantaged families across neighborhoods.

I next ask whether these differences in neighborhood choice can explain the disparities between families in Figure 2. Columns 2 and 3 report regressions of child income on predicted income, with and without neighborhood fixed effects. By construction, the coefficient without fixed effects is equal to one. Including neighborhood fixed effects restricts comparisons to

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

[24]I use that the variance in predicted incomes is $V(\hat{Y}) = 125$ and that the neighborhood fixed effects explain 30 percent of this variance: $V(N)/V(\hat{Y}) = 0.3$. Then it follows that $SD(N) = \sqrt{0.3 \cdot 125} = 6.1$.

children growing up in the same neighborhood. If predicted income gaps are driven by differences between neighborhoods, the coefficient should decline. However, the coefficient falls only slightly to 0.94. This implies that the income disparities by family background are almost equally strong within neighborhoods as between neighborhoods.[25]

Neighborhoods also explain little income variation that was left unexplained by the family background variables. Adding the neighborhood fixed effects increases the adjusted R^2 only from 16.5 percent to 17.1 percent.

In summary, although families sort into neighborhoods, neighborhoods explain only a small portion of income differences linked to family background and contribute little beyond what is already captured by the observed characteristics. This suggests that the overall role of neighborhoods in driving income inequality is modest.[26]

Table 3: Predicted and observed income differences between neighborhoods

	Predicted income	Income	Income
	(1)	(2)	(3)
Predicted income	-	1.003	0.943
		(0.004)	(0.005)
N	333,792	333,792	333,792
Adjusted R^2	0.298	0.165	0.171
Neighborhood fixed effects	x		x

Notes: Column 1 reports the adjusted R^2 of a regression of children's predicted income rank on neigbhorhood fixed effects. Columns 2 and 3 report results from separate regressions of a child's income rank on its predicted value, with and without neighborhood fixed effects. The predicted incomes are based on the gradient-boosted decision trees reported in figure 2. The sample correspond to all children from the test data with an available neighborhood identifier (97%). Standard errors, shown in parentheses, are clustered at the neighborhood level in column 3.

6.2 The Post-birth Environment

One plausible explanation for the limited role of neighborhoods is that much of family (dis)advantage is already realized during pregnancy or shortly thereafter. For example,

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

[26]Of course, this does not imply that assigning children to different neighborhoods will not affect them.

children born into advantaged families may benefit from favorable genetic endowments or higher quality prenatal care, both of which can shape long-term income trajectories. To assess the relative importance of pre-birth versus post-birth influences in explaining the disparities shown in Figure 2, this section examines the causal effect of being assigned shortly after birth to a family associated with a 1 rank higher predicted income.

To answer this question, I use a sample of 4,935 international adoptees born between 1980 and 1989 and who arrived in the Netherlands within six months of birth.[27] 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. In addition, a major advantage of using adoptees from the same cohorts and using exactly the same variables is that I can easily compare the results to the main result in Figure 2.

The analysis mimics previous studies using international adoptees (e.g. Sacerdote (2007), Holmlund et al. (2011), Fagereng et al. (2021)). This section extends previous results by characterizing family background with a multidimensional measure - the income prediction, which is based on many family background variables - and by focusing on children's long-run income ranks. Despite its central role in intergenerational mobility analyses, this dimension has been overlooked in studies using international adoptees.[28]

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 place-

[27] Although the Netherlands lacks an adoption register, Statistics Netherlands developed a reliable method to identify adoptees. They sent a survey to a random subset of all plausible adoptees to verify their method. Overall, 97.7 percent of respondents in my sample confirmed they were adopted ($n = 778$). Nevertheless, a minor fraction of plausible adoptees may not be adopted. This may induce a small upward bias in the estimate in Columns 1 to 4 of Table 4.

[28] Sacerdote (2007) also considers international adoptees and examines child incomes. However, as the author acknowledges, the income measure is imperfect, complicating comparability. See Black et al. (2020) for an analysis with domestic adoptees.

ment, as prioritizing specific characteristics would have significantly increased already long waiting times.[29] Second, the relatively large sample allows me to include controls for gender, age at migration, and fully interacted fixed effects for the country and year of adoption, which are all observable characteristics of the child at the time of adoption. Adding such controls does not alter the estimates, indicating that selective placements based on these observable characteristics are of limited empirical importance.

Column 1 of Table 4 shows that being raised in a family that is associated with a 1 rank higher income for own-birth children increases the income rank of adoptees by only 0.28. Columns 2 to 4 show that this result is robust towards the inclusion of controls. Assuming no selection bias and generalizability towards the broader population, this estimate suggests that around 30% of the disparities in Figure 2 are shaped by the post-birth environment.

Although adoptees and their adoptive parents are clearly not representative of the broader population, I also offer two reasons why external validity concerns may not be overly severe. First, column 5 in Table 4 shows that the association between realized and predicted income stays close to one for own-birth children in families with at least one adopted child. This indicates that differences in the predictability of income between adoptees and own-birth children are not driven by fundamental differences between families with and without adoptees. Second, while there are no highly disadvantaged adoptive families, Table A4 reveals substantial variation in the characteristics of adoptive families, spanning a broad range of the general population. Nevertheless, concerns such as differential treatments of adopted children by parents or others still apply.

To conclude, the results show that growing up from birth in an advantaged family does improve income prospects, but considerably less than for own-birth children. This finding underscores the importance of pre-birth factors as drivers of intergenerational persistence.

[29] 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`.

Table 4: The effect of family background on income: regression results with adoptees

	Household income rank				
	(1)	(2)	(3)	(4)	(5)
Predicted income rank	0.279	0.282	0.285	0.282	0.872
	(0.028)	(0.029)	(0.031)	(0.027)	(0.047)
Controls		x	x	x	
Country of Origin FE			x	x	
Year of Adoption FE				x	
N	4,935	4,935	4,935	4,935	3,802
Sample	Adoptees	Adoptees	Adoptees	Adoptees	Families with at least one adopted child

Notes: Columns 1 to 4 show results from separate regressions of adopted children's household income ranks on their predicted income ranks based on the family background variables. Column 5 shows results of the same regression, now applied to own-birth children in families with at least one adopted child. All families with adopted children were excluded from the training data. 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. (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$)

7 Conclusion

Understanding the sources of income inequality is central to debates about the need for redistributive or opportunity-equalizing policies (Almås et al. (2024)). While individuals often accept income differences that reflect effort or preferences, disparities arising from external circumstances such as parental background are widely viewed as illegitimate (Alesina et al. (2018), Almås et al. (2020)). A careful assessment of the role of family background in shaping children's economic outcomes therefore provides essential information for policymakers and the broader public in evaluating the fairness of current income inequality.

This paper contributes by studying the importance of family background for children's economic success using an exceptionally detailed set of family background information. 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. I go beyond aggregate statistics and provide the entire distribution of children's expected incomes alongside the corresponding family background characteristics. Additionally, I provide new insights on the roles of neighborhoods and pre-

birth versus post-birth factors in driving the observed inequalities. These results shed new light on the extent to which family background shapes children's economic success.

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

References

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

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

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

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

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

Almås, Ingvild, Alexander W Cappelen, Erik Ø Sørensen, and Bertil Tungodden. 2024. "Attitudes to Inequality: Preferences and Beliefs." *Oxford Open Economics* 3 (Supplement_1): i64–i79.

Almås, Ingvild, Alexander W. Cappelen, and Bertil Tungodden. 2020. "Cutthroat Capitalism versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking than Scandinavians?" *Journal of Political Economy*.

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.

Besemer, Sytske, Shaikh I. Ahmad, Stephen P. Hinshaw, and David P. Farring-

ton. 2017. "A Systematic Review and Meta-Analysis of the Intergenerational Transmission of Criminal Behavior." *Aggression and Violent Behavior* 37 161–178.

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.

Black, Sandra E, Paul J Devereux, Petter Lundborg, and Kaveh Majlesi. 2020. "Poor Little Rich Kids? The Role of Nature versus Nurture in Wealth and Other Economic Outcomes and Behaviours." *The Review of Economic Studies* 87 (4): 1683–1725.

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`.

Boustan, Leah, Mathias Fjællegaard Jensen, Ran Abramitzky et al. 2025. "Intergenerational Mobility of Immigrants in 15 Destination Countries." NBER Working Paper No. 33558.

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

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

Brunori, Paolo, Francisco H.G. Ferreira, 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.

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

Corak, Miles. 2020. "The Canadian Geography of Intergenerational Income Mobility." *The*

Economic Journal 130 (631): 2134–2174.

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

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

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

Eriksson, Karin Hederos, Randi Hjalmarsson, Matthew J. Lindquist, and Anna Sandberg. 2016. "The Importance of Family Background and Neighborhood Effects as Determinants of Crime." *Journal of Population Economics* 29 (1): 219–262.

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

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

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

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

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

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

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

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

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

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

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

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

Manduca, Robert, Maximilian Hell, Adrian Adermon et al. 2024. "Measuring Absolute Income Mobility: Lessons from North America and Europe." *American Economic Journal: Applied Economics* 16 (2): 1–30.

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.

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

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

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

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

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

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

Sacerdote, Bruce. 2007. "How Large Are the Effects from Changes in Family Environment? A Study of Korean American Adoptees." *The Quarterly Journal of Economics* 122 (1): 119–157.

Santavirta, Torsten, and Jan Stuhler. 2024. "Name-Based Estimators of Intergenerational 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: supplementary results

Table A1: Descriptive statistics for the income analysis sample

	Mean	SD	Mean	SD	% missing
Characteristics children					
Year of birth	1984.6	2.9			0
Male	0.51	0.50			0
Family size	2.7	1.3			0
Household income	102156	65404			0
Second generation migrant	0.15	0.36			0
Third generation migrant	0.06	0.23			0
Family characteristics: measured at the household level					
Household income rank	0.50	0.29			0.009
Primary income share	0.794	0.268			0.011
Highest education	12.937	3.637			0.358
Total wealth rank	0.50	0.29			0.008
Bank and savings balances	52,249	180,945			0.008
Bonds and shares	36,704	347,226			0.008
House value	309,747	379,964			0.008
Entrepreneurial assets	15,028	132,290			0.008
Other real estate	30,253	277,509			0.008
Substantial interest	65,601	1,235,768			0.008
Other assets	6,091	111,069			0.008
Total debt	159,239	374,080			0.007
Mortgage debt	134,709	190,726			0.008
Relationship status of household head(s) of child at age 15:					
Registered partners	0.824	0.381			0.023
Non-registered partners	0.037	0.19			0.023
Single parent	0.126	0.332			0.023
Other	0.012	0.11			0.023
Other family characteristics	**Father**		**Mother**		
Personal income	68,129	51,443	29,157	21,734	0.108
Personal earnings	83,082	61,812	33,161	26,958	0.180
Most important source of income					
Employment	0.669	0.416	0.536	0.433	0.055
Bonds or shares	0.043	0.179	0.012	0.090	0.055
Entrepreneurship	0.116	0.288	0.066	0.218	0.055
Substantial interest	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
Other social security	0.004	0.049	0.007	0.062	0.055
Disability insurance transfers	0.079	0.237	0.064	0.212	0.055
Pension	0.023	0.109	0.037	0.147	0.055
Other	0.014	0.087	0.185	0.338	0.055
Type of housing					
Own house	0.745	0.409	0.7	0.428	0.066
Rental	0.053	0.19	0.104	0.259	0.066
Subsidized rental	0.2	0.356	0.195	0.338	0.066
Years of education	12.785	3.832	11.934	3.666	0.53
Average hourly wage	32.005	26.927	20.691	18.097	0.315
Most important sector of employment	In 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	0.156	0.006	0.079	0.014
Suspected of other crime	0.042	0.2	0.012	0.11	0.014
Total health costs	2,700	7,153	2,626	8,212	0.014
General practitioner costs	174	143	197	155	0.063
Mental health care costs	234	3,541	321	3,948	0.063
Hospital care costs	1,830	6,723	1,692	5,013	0.063
Pharmaceutical care costs	527	2,230	542	2,084	0.063
Dental care costs	46	303	44	299	0.063
Age at first birth	29.285	5.546	26.952	4.394	0
Family size	4.14	2.365	4.044	2.299	0.218
Birth order	2.481	1.777	2.502	1.8	0.218
Father/mother identified	0.025	0.157	0.002	0.049	0
Father/mother dead	0.008	0.086	0.004	0.065	0.019
Father/mother present in household	0.857	0.35	0.962	0.191	0.037
Migration background	In 8 categories				0.315
Migration background grandfather	In 8 categories				0.315
Migration background grandmother	In 8 categories				0.315
Extended family outcomes					
Average income rank	0.496	0.222	0.495	0.224	0.246
Average education	12.61	3.155	12.732	3.103	0.42
Average wealth rank	0.514	0.226	0.511	0.227	0.239
Average health expenditures	2717	5537	2564	5370	0.231
% of siblings suspected of any crime	0.043	0.142	0.048	0.153	0.231

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

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

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

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

In addition, I identify the primary sources of personal income, classified into 10 categories.[30] Drawing on all yearly observations used in constructing the lifetime personal income measure, I first compute the most important source of income in each of those years. I then compute the fraction of years in which each category served as the main source of income.

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

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

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

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

Education. Parents' years of education are based on the conversion table in Appendix D1. Table A1 indicates that parental education information is absent for about 50 percent of the sample. This gap exists because Statistics Netherlands initiated systematic education data collection only in the late 1980s. Prior educational records are mainly sourced from large-scale surveys frequently administered by Statistics Netherlands and are also obtained indirectly from other government bodies, including the unemployment agency.

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

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

Crime. As explained in section 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.

Migration background. I have information on the country of origin of all identified parents and grandparents. I distinguish eight regions: the Netherlands, Morocco, Turkey, Surinam, Dutch Antilles, Western Europe, Eastern Europe, and others.

Extended family outcomes. For each parent separately, I determine the mean years

of education, household income rank, wealth rank, and annual health expenditures across all their siblings. Additionally, I calculate the fraction of these siblings who have been suspected of committing a crime.

Table A2: Predicting child income using smaller samples

Share of core sample	Test data sample size	R^2	0.025% lower bound	97.5% upper bound
(1)	(2)	(3)	(4)	(5)
0.01	3,406	0.139	0.118	0.163
0.02	6,812	0.148	0.132	0.166
0.05	17,031	0.153	0.143	0.162
0.1	34,061	0.159	0.152	0.166
0.2	68,122	0.159	0.154	0.164
0.4	136,243	0.164	0.160	0.167
0.6	204,365	0.164	0.160	0.166
0.8	272,486	0.163	0.161	0.166

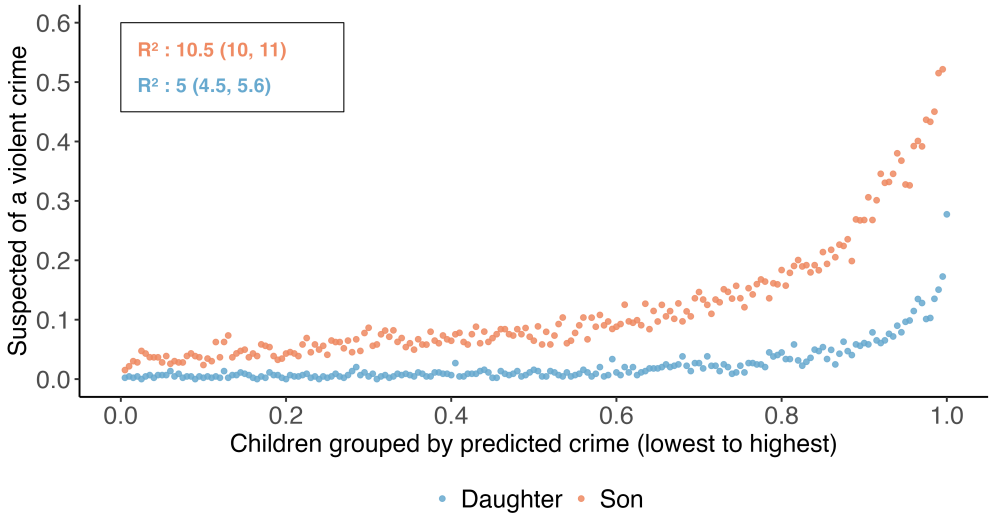
Notes: This table presents estimates of explanatory power for gradient-boosted decision trees that include all explanatory variables (as in Figure 2), using smaller samples. Column 1 reports the share of the core sample that is used for the analysis. Column 2 reports the sample size of the test-data. 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 A3: Predicting child income: varying years and ages of income measurement

	R^2	0.025% lower bound	97.5% upper bound
Years of income	A. Varying years of income measurement		
1	0.138	0.134	0.142
2	0.145	0.142	0.150
3	0.151	0.147	0.156
4	0.153	0.149	0.157
5	0.157	0.153	0.162
6	0.158	0.154	0.162
7	0.162	0.158	0.166
8	0.161	0.157	0.166
9	0.165	0.161	0.170
All	0.170	0.166	0.174
All > age 32	0.166	0.162	0.170
Age child	B. Varying ages of income measurement		
30-33	0.129	0.125	0.133
34-37	0.154	0.150	0.159
38-41	0.153	0.149	0.158

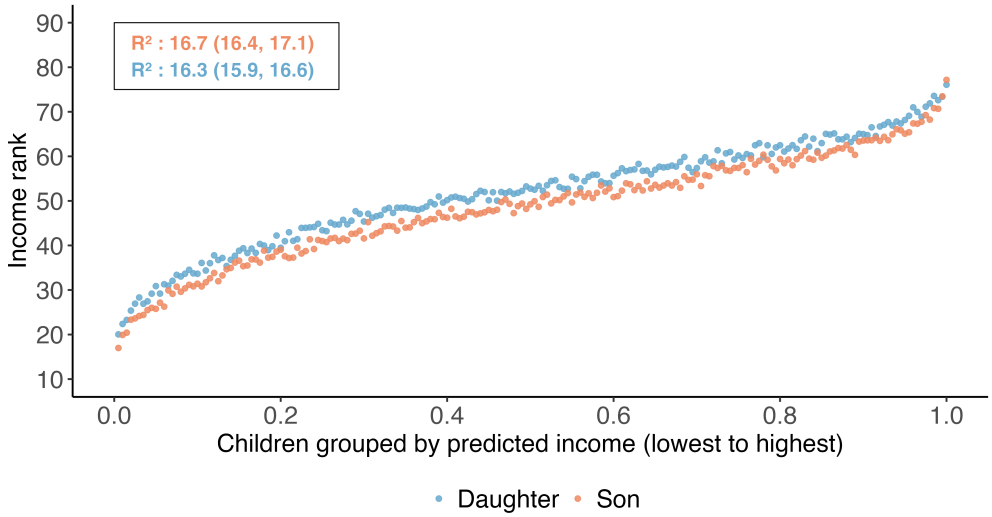
Notes: Each row presents the R^2 and corresponding 95% lower and upper bound for gradient-boosted decision trees that include all explanatory variables to predict child income (as in Section 5). The analysis sample consists of all 330,018 children born in 1980 and 1981 for whom I observe all incomes between ages 30 and 41. Each model is trained on the same randomly selected 80% of this sample, and evaluated on the remaining 20%. Panel A varies the number of years of income data used to construct the child income rank. The one-but-last row in panel A uses all income observations, as in the main results. The last row uses all income data above age 32. Panel B uses four years of income data, but varies the ages at which income is measured. Confidence intervals for the R^2 are bootstrapped from the test-data using 599 draws.

Figure A1: Predicting children's violent crime by gender



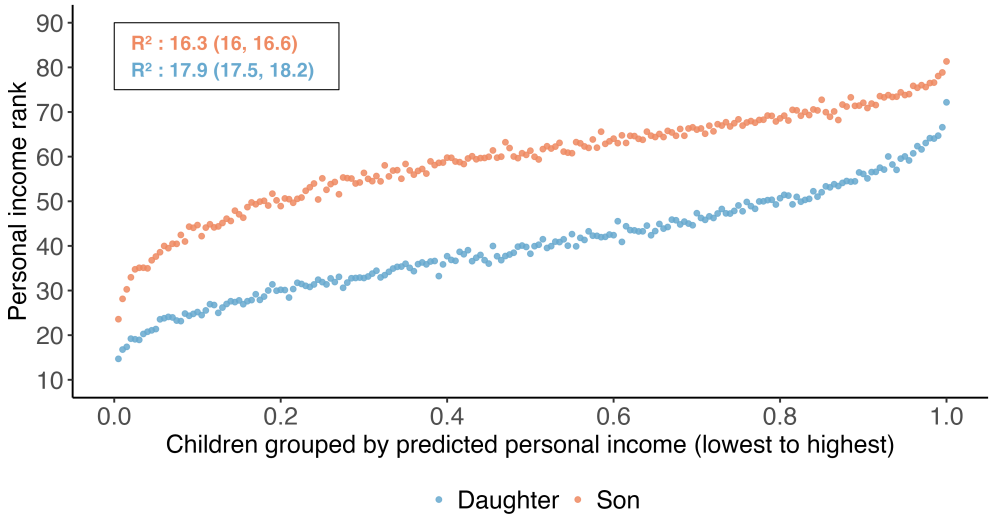
Notes: this figure presents a scatter plot of predicted crime for 92,725 sons and 89,051 daughters separately. Crime is measured as an indicator that equals 1 if a child was suspected of a violent crime between the ages of 20 to 33. The graphs are constructed using the same steps as in Figure 2. Confidence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

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



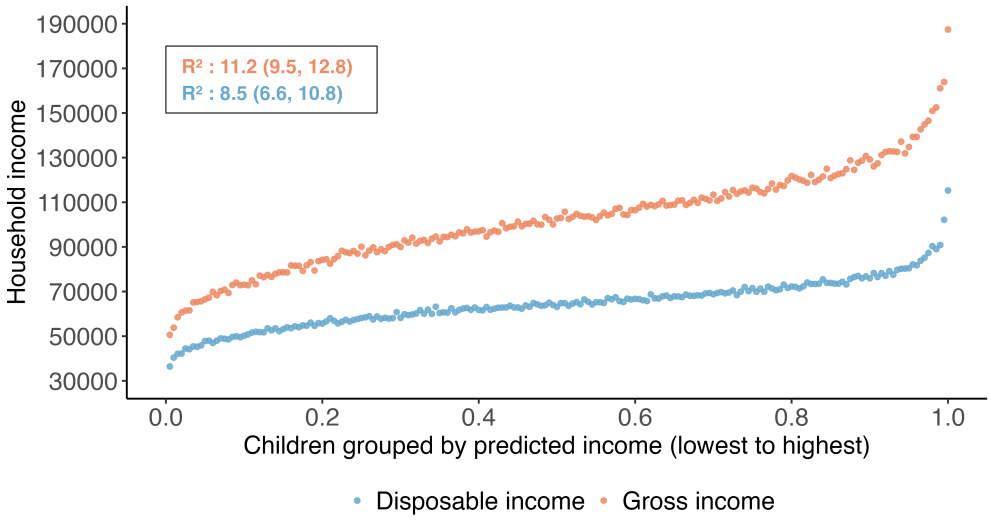
Notes: this figure presents binscatter plots of sons' and daughters' household income ranks for 173,652 sons and 166,957 daughters in the test data, who are sorted into bins based on their predicted income rank. Predictions are generated using the same predictive model and explanatory variables as in Section 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. Confidence intervals for the R^2 are bootstrapped from the test data using 599 samples and are reported in brackets

Figure A3: Predicting children's personal income by gender



Notes: this figure presents binscatter plots of sons' and daughters' personal income ranks for 172,976 sons and 164,990 daughters in the test data, who are sorted into bins based on their predicted income rank. The graphs are constructed using the same steps as in Figure 2, 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

Figure A4: Predicting (disposable) household income levels



Notes: this figure presents binscatter plots of children's gross household income and disposable household income for 340,608 children in the test data, who are sorted into bins based on their predicted income rank. The graphs are constructed using the same steps and sample as in Figure 2, applied to children's gross household income and disposable household 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

Table A4: Descriptive statistics for international adoptees and their parents

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Income rank (y)	35.35	37.74	38.91	38.66	42.26	43.06	40.96	42.04	42.97	43.21
Predicted income rank ($\hat{y}$)	38.24	45.72	49.21	51.94	54.4	56.67	58.79	61.04	63.98	69.44
	Characteristics Adoptive Parents									
Parental income rank	19.27	29.86	37.26	43.54	53.65	59.94	68.24	76.43	84.97	93.48
Parental wealth rank	32.16	46.48	54.18	58.01	62.94	66.46	68.92	70.05	72.64	81.42
Highest education parents	11.21	11.88	13.01	13.39	14.4	14.66	14.76	15.4	15.86	16.45
Father suspected of crime	0.10	0.04	0.04	0.03	0.04	0.02	0.03	0.02	0.02	0.02
Health expenditures parents	4,499	3,788	3,475	2,890	3,180	2,592	2,725	2,645	2,630	2,266
Extended family income rank	38.03	44.61	47.72	50.53	51.63	55.81	58.75	60.54	61.86	71.04
N	493	494	493	494	493	494	493	494	493	494

Notes: Each column shows descriptive statistics for a group of international adoptees from the same predicted income bin. The predicted income bins are constructed by predicting the income ranks of all adoptees using the model with all explanatory variables (as in Figure 2), ranking them from low to high, and sorting them into ten equally sized bins according to their position in the predicted income distribution. All cells are averages.

Appendix B: intergenerational mobility estimates

Additional results. Given that my baseline intergenerational mobility estimate differs from other estimates in the Netherlands, I provide additional estimates here that are commonly reported in the literature. These can be used by other researchers that wish to make cross-country comparisons. Below, I also present a sensitivity analysis and elaborate on why my estimates differ from prior estimates.

Table B1 reports the rank-rank correlation as well as the Intergenerational Income Elasticity (IGE) using logs of household income instead of ranks in columns 1 and 2. These are, coincidentally, equal up to the second digit. Columns 3 and 4 report results for sons and daughters separately and rely on children's personal income ranks instead of household income ranks. These estimates are very similar between genders and somewhat lower than the rank-rank correlation based on household income. Finally, column 5 reports the sibling correlation in income. This estimate suggests that about 30% of all variation in income ranks is driven by factors shared between siblings.

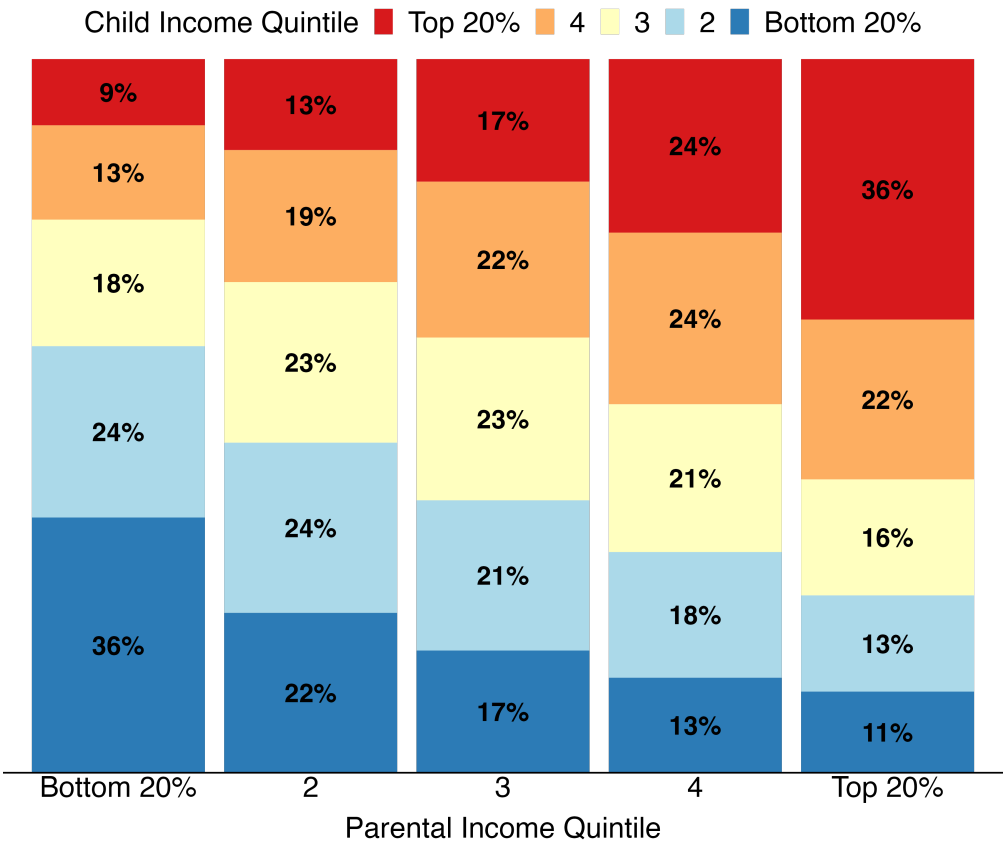
Figure B1 reports a 5×5 transition matrix. This table can be used to compare upward or downward mobility estimates across countries.

Table B1: Intergenerational mobility estimates

	Rank rank correlation	IGE	Personal income rank (daughters)	Personal income rank (sons)	Sibling correlation
	(1)	(2)	(3)	(4)	(5)
Coefficient	0.323	0.324	0.288	0.290	-
	(0.001)	(0.001)	(0.001)	(0.001)	
N	1,702,355	1,702,355	864,064	825,170	1,702,355
Adjusted R^2	0.105	0.091	0.093	0.095	0.308

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. Column (5) reports the sibling correlation. This is estimated by the adjusted R^2 of a regression of child income on sibling fixed effects and the child's year of birth. The sample includes the core analysis sample (Table A1) excluding observations with missing parental income (0.9 percent). Standard errors are in parentheses.

Figure B1: Transition matrix



Notes: This figure presents the transition matrix of child income conditional on parental income quintile. Each bar represents the distribution of child income quintiles for children whose parents fall in the corresponding parental income quintile on the x-axis. The segments within each bar show the share of children reaching each income quintile, as indicated by the color legend. The sample includes the core analysis sample (Table A1) excluding observations with missing parental income (0.9 percent).

Sensitivity. Next, I 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.

Table B2 reports mobility estimates using varying years of income information of parents. I focus on all children for whom both the father and the mother have at least 9 observable income observations. 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. This suggests that attenuation bias is unlikely to be an issue.

Table B3 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 4 years for each of the specifications. The estimates are very similar, regardless of when parental income is measured.

Table B4 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 5 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.

Table B2: Intergenerational mobility estimates: varying years of parental income

Years of income	1	2	3	4	5	6	7	8	9
Coefficient	0.297	0.304	0.311	0.316	0.320	0.323	0.325	0.327	0.329
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,098,025	1,098,025	1,098,025	1,098,025	1,098,025	1,098,025	1,098,025	1,098,025	1,098,025
R^2	0.082	0.086	0.090	0.092	0.095	0.096	0.098	0.099	0.100

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 parental 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 for whom at least 9 paternal *and* 9 maternal incomes are available.

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

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

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

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

	(1)	(2)	(3)
Coefficient	0.274	0.304	0.308
	(0.002)	(0.002)	(0.002)
Age child	30-33	34-37	38-41
N	326,388	326,388	326,388

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 for whom all incomes between ages 30 and 41 are available. Standard errors are reported in parentheses.

Comparison with other studies. There are three recent estimates of the rank-rank correlation in the Netherlands.

Most closely related is Van Elk et al. (2024). They study intergenerational mobility differences among migrants and natives, and use the same data as in this paper. While in the main paper they focus on disposable household income, in the Appendix, they report a rank-rank correlation of 0.22 that corresponds to gross household income. There are four main differences between our approaches. Below, I describe these differences and quantify their importance in Table B5 step by step.

The core analysis sample in this paper includes all children born between 1980 and 1989, excluding only 3.4% of children with missing income observations. Van Elk et al. consider children born between 1983 and 1988. Column 1 of Table B5 replicates the rank-rank correlation for children born in these years. For these cohorts, I find a similarly large rank-rank correlation of 0.33. Starting from this baseline estimate, I change my measurement approach so as to align with Van Elk et al.

First, Van Elk et al. drop all children who do not live independently in 2003 and who do live independently in 2017 to 2019, whereas I do not make such sample restrictions. Dropping these individuals results in a 23 percent smaller sample and reduces the rank-rank correlation by 0.023 (columns 2 and 3).

Second, Van Elk et al. measure child income from 2017 to 2019, when children are aged 29 to 36. I average income over all available observations from age 30 onward and up to 2023. Implementing their age at measurement further reduces the rank-rank correlation by 0.021 (column 4).

Third, Van Elk et al. measure parental income from 2003 to 2005. I measure parental income over all available observations from 2003 and up to age 60. On average, that corresponds to 12 observations for fathers and 14 observations for mothers. Implementing their parental age at measurement further reduces the estimate by 0.029 (column 5).

Fourth, Van Elk et al. define parents as the head of the child's household in 2003 and his or her partner. Parental income is then defined as the income of this household head and his or her partner between 2003 and 2005. Instead, I define parents based on legal relationships extracted from birth certificates. Following Chetty et al. (2014), parental income is then defined as the average of the household income of the father and the mother. Our parental income concepts align when the child, father, and mother live together between 2003 and 2005. However, when at least one of the legal parents is not present in the household in

these years, our definitions differ. Implementing their measure of parental income reduces the estimate by 0.055, resulting in an estimate that is very close to their main estimate (column 6).

This drop is relatively large because the legal father or mother is absent from the child's household in 2003 in 28 percent of cases. For these children, the income of the legal parents is considerably more predictive than that of their household heads.[31]

Table B5: Comparison with Van Elk et al. (2024)

	(1)	(2)	(3)	(4)	(5)	6)
Coefficient	0.329	0.313	0.306	0.285	0.256	0.201
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	1,016,358	883,471	779,159	779,159	778,998	778,998
Adjustments						
Child born between 1983-1988	x	x	x	x	x	x
Child living independently in 2003		x	x	x	x	x
Child living independently in 2017 to 2019			x	x	x	x
Child income measured in 2017 to 2019				x	x	x
Parental income measured in 2003 to 2005					x	x
Parental income based on household head						x

Notes: each column presents an estimate of the rank-rank correlation, using different variable definitions and sample selections. The specification in Column 1 uses the same variable definitions as in the main text (Table B1), focusing exclusively on children born between 1983 and 1988. The subsequent columns report results using a different sample selection or different variable definitions. These differences are further explained in the main text above.

There are also two estimates of the rank-rank correlation which are based on different data. First, Boustan et al. (2025) compare intergenerational mobility among migrants and natives in 15 destination countries, including the Netherlands. While the children's incomes are based on the same population-wide administrative data, the parents' incomes in their study are based on a random sample of administrative data from before 2003 (in Dutch: the 'IPO'). This random sample contains incomes in 1981, 1985, and annually from 1989 for about 3.3% of the population. Boustan et al. report intergenerational mobility estimates of 0.24 and 0.22 for sons and daughters born between 1982 and 1987 (See Table C.9.23). There are three main differences with their approach: (i) they use children's personal income (in 2018 and 2019), whereas I use household income (measured above age 30 and up to 2023), (ii) they use the sum of parents' personal incomes instead of parental household income, and (iii) they measure parental income from 1998 to 2004, whereas I measure parental income from 2003 and up to age 60.

[31] Using the sample of children for whom at least one of the legal parents is not present in the household, I find a rank-rank correlation of 0.32 when using the legal parents' incomes. This drops to 0.05 when using the household head and his/her partner's income.

The personal income measure of Statistics Netherlands excludes not only the partner's income but also income components from joint tax statements that cannot be attributed to specific individuals. These include income from wealth and allowances allocated based on household-level income, such as child and rental allowances. Consequently, the sum of parents' personal incomes does not match the household income measure provided by Statistics Netherlands, which I employ in this study, even for cohabiting parents.

I do not have access to the survey, precluding a direct comparison with my results. However, in Table B6, I try to mimic their analysis as closely as possible, using the population wide administrative data. I begin by restricting my sample to children born between 1982 and 1987 and estimate the baseline rank–rank specification, which yields a correlation of 0.33.

In column 2, I revise the parental income measure to the sum of both parents' personal incomes from 2003 to 2009. While I cannot observe incomes prior to 2003, this at least aligns the number of years over which parental income is measured.[32] This reduces the rank–rank correlation to 0.29. I then replace my original outcome with the child's personal income rank, based on income measured in 2018 and 2019. This further reduces the estimate to 0.256 (column 3), which is quite close to their estimate. Remaining differences may reflect discrepancies between survey and administrative data, for instance due to missing income information for non-cohabiting parents in the survey.

Lastly, Manduca et al. (2024) study trends in absolute mobility across multiple countries. While their main goal is not to quantify relative intergenerational mobility, they also report rank-rank correlations for the Netherlands from 0.23 in the 1974 cohort to 0.16 for the 1984 cohort. They use a very similar approach as Boustan et al. They also link children's incomes from the population wide administrative data to parental income from the representative survey, and also rely on personal income measures for children and parents. The main difference with Manduca et al. is that Boustan et al. measure parental and child income in only one year (the closest observation to age 30 for both generations). As shown in Table B6 column 4, using only one income observation for parents and children and measuring child income at age 30 further reduces the estimate to 0.22.

Since I do not observe parents' incomes before 2003, I cannot assess the impact of also measuring parental income at age 30. However, Table B4 shows that results attenuate somewhat when measuring child incomes in the early 30s, suggesting that individuals may not be on their long-term income trajectory at that age. A similar bias may occur when measuring parental income at this relatively young age.

[32]Estimates are stable across different years of income measurement between 2003 and 2013 (Table B3). This makes it likely that estimates are also similar when parental income is measured between 1998 and 2004.

Table B6: Comparison with Boustan et al. (2025) and Manduca et al. (2024).

	(1)	(2)	(3)	(4)
Coefficient	0.327	0.292	0.256	0.229
	(0.001)	(0.001)	(0.001)	(0.001)
N	986,125	986,125	986,125	986,125
Adjustments				
Child born between 1982 and 1987	x	x	x	x
Using personal income of parents		x	x	x
Using personal income of child in 2018 and 2019			x	x
Using one income observation for parents (in 2003) and children (at age 30)				x

Notes: each column presents an estimate of the rank-rank correlation, using different variable definitions and sample selections. The specification in Column 1 uses the same variable definitions as in the main text (Table B1), focusing exclusively on children born between 1982 and 1987. The subsequent columns report results using a different sample selection or different variable definitions. These differences are further explained in the main text above.

Appendix C: Measuring variable importance

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

Shapley values originate from cooperative game theory (Shapley (1953)). In this framework, a coalition of agents $j \in S$ produces an output $\nu(S)$. The Shapley value for agent $i \in S$ represents its average marginal contribution to the output $\nu(s)$ across all possible coalitions $s \subseteq S \backslash i$. This concept directly applies to prediction models, where the output $f(x_1, ..., x_k)$ is generated from a set of variables $x_j \in X^j$, with $j \in \{1, ..., k\}$. In this context, Shapley values represent the average marginal contribution of each variable to a prediction, calculated by averaging over all possible subsets of included variables.[33]

Lundberg and Lee (2017) show that Shapley values are the only measures of variable importance that preserve important properties from cooperative game theory.[34] While exact Shapley values are computationally infeasible for most models due to the need to sum over all feature subsets (an NP-hard problem), recent algorithms can compute exact Shapley values for tree-based models in short time periods (Lundberg et al. (2020)).

Using this algorithm, I compute Shapley values for the gradient-boosted decision tree model used in the main results (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. Of all explanatory variables, only the indicators for whether the father or mother is identified in the data provide no contribution to the predictions. Figure C1 presents a boxplot of Shapley values for the 30 variables with the highest average absolute Shapley values, ranked from highest to lowest.

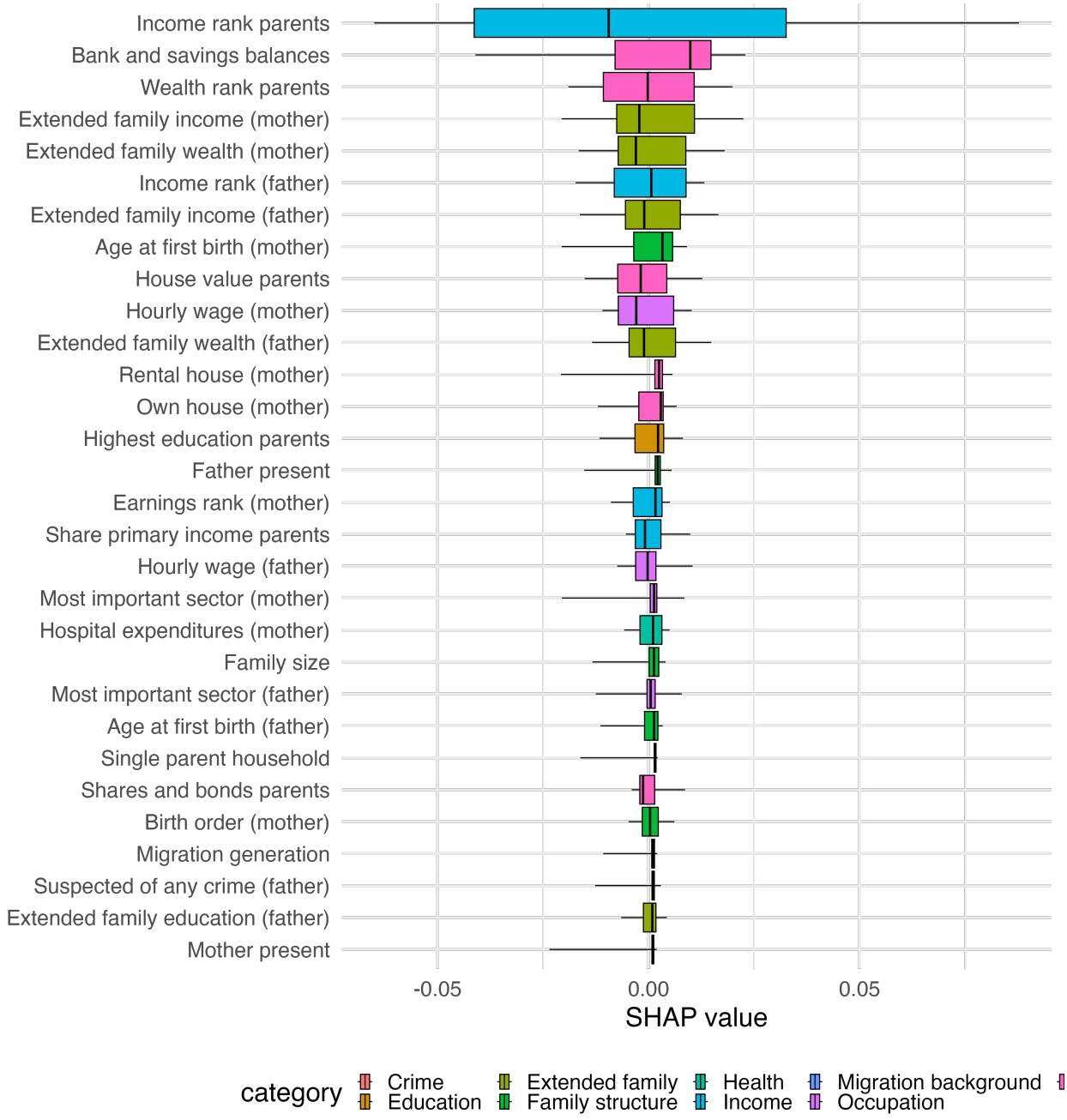
To illustrate, consider the boxplot for the household income rank: the 2.5th percentile is -0.07, indicating that for 2.5 percent of the children, parental income reduces the prediction by at least 0.07 ranks compared to the average prediction of 0.50. The 75th percentile is 0.03, meaning that for 25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

Figure C1 shows that nine of the ten 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.

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

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

Figure C1: Measuring variable importance using Shapley values



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

Appendix D: a 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.

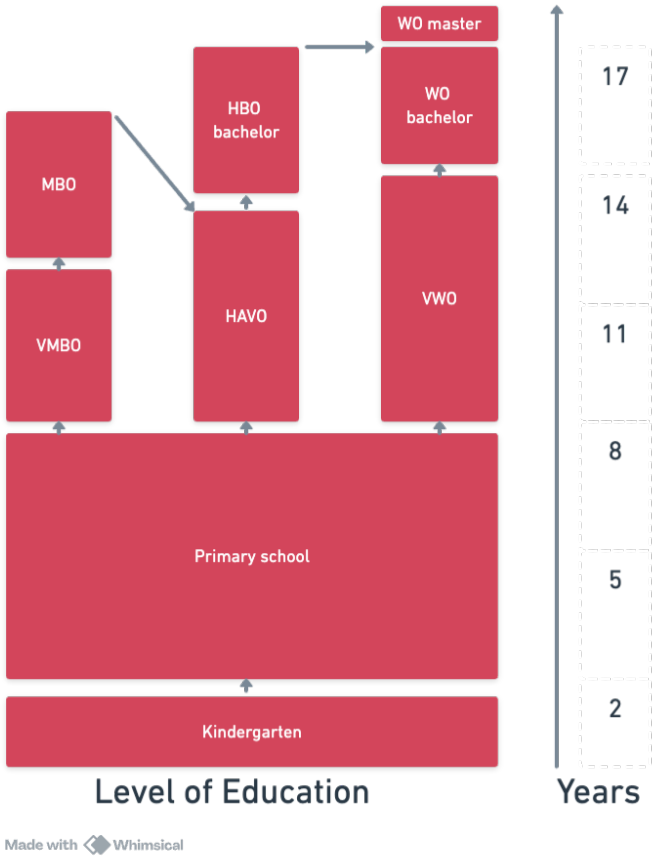


Figure D1: The Dutch Educational System

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