

Who Gets Ahead? Measuring Income Gaps across Family Backgrounds

Sander de Vries*

September 4, 2025

Abstract

Measuring income gaps across family backgrounds is crucial for evaluating inequality and informing policy. I develop a multidimensional approach and apply it to Dutch administrative data with exceptionally rich information on parents and extended family. This approach reveals strikingly larger income gaps than standard intergenerational mobility measures, particularly among the most disadvantaged children. A novel decomposition shows that income gaps are largest along parental and extended family income and wealth, though other characteristics matter greatly for specific subgroups. Neighborhoods contribute modestly. Evidence from adoptees suggests that pre-birth factors are an important driver of the observed disparities.

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, Stanislav Avdeev, and conference and seminar participants in Amsterdam, Utrecht, The Hague, Tokyo, London, Canazei, Colchester, Antwerp, and Sydney. Earlier versions of this paper were circulated under the title ‘Measuring Family (Dis)Advantage: Lessons from Detailed Parental Information’. 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

Concerns about inequality often focus on the role of family background in shaping children’s economic success. Low income prospects for children from disadvantaged families are widely seen as a failure of equal opportunity and as a justification for policy intervention (Roemer and Trannoy (2016), Alesina et al. (2018)). Yet, despite its importance, measuring the size and nature of such income gaps remains challenging.

Recent administrative data have allowed researchers to map in detail how parental and child incomes are related (e.g., Chetty et al. (2014), Deutscher and Mazumder (2020), Kenedi and Sirugue (2023)). These studies reveal which children have the lowest and highest expected incomes and their chances of moving up or down the income distribution. Yet even among low-income families, some children face especially poor prospects when the family is also disadvantaged in other dimensions such as education or family structure. Analyses based on parental income alone miss these disparities. While measures such as sibling correlations or inequality of opportunity estimates aim to capture such broader influences, they typically summarize it in a single statistic and therefore conceal how large the income gaps are between specific types of families.¹ As a result, two important questions remain unresolved: how large are income gaps across broader family background characteristics, and what distinguishes the families where children have systematically low or high incomes?

This paper sheds light on these questions by addressing two main challenges. The first is that a rich characterization of family background requires linked child-parent data with information across many dimensions. I draw on Dutch administrative data that link the long-run incomes of 1.7 million children to detailed information on their parents and extended family, including income, wealth, occupation, education, crime, health, migration background, and family structure. This makes it possible to examine disparities between

¹Deutscher and Mazumder (2023) distinguish between ‘global’ measures, which summarize the overall role of family background, and ‘local’ measures, which provide distributional detail. They also separate ‘narrow’ measures, based on a single parental characteristic, from ‘broad’ measures that capture the broader family background. They show that no existing study offers a broad but local analysis.

families in much greater detail than previous work.

The second challenge is to translate this high-dimensional information into interpretable measures that reveal fine-grained distributional patterns. I use all available family information in a flexible prediction model of child income and employ these predictions to describe both the distribution of expected income and children’s probabilities of reaching different parts of the income distribution. In addition, I adapt a new method based on Shapley values to describe how different family background characteristics contribute to children’s expected income ranks (Lundberg et al. (2020)). Unlike traditional variance decomposition methods, this method reveals family background characteristics that are important for some individuals even if their aggregate contribution is modest.

As a starting point, I examine the relationship between child and parental income, following the standard intergenerational mobility approach. I estimate a rank-rank correlation for the Netherlands of 0.32 ($R^2 = 10.5$ percent), which is high compared to other OECD countries.² This shows that parental income alone already predicts substantial income gaps.

I then show that focusing on parental income alone severely understates income disparities across family backgrounds, especially for the most disadvantaged children. Incorporating all family information raises explanatory power by 58 percent, with these family characteristics jointly accounting for 16.6 percent of income variation. The difference between the two approaches is greatest at the bottom of the distribution. For instance, the 0.5 percent of children with the lowest expected incomes based on parental income alone have an average observed rank of 31 (out of 100). With all family information, this falls to 18. These children’s probability of reaching the top income quintile is less than two percent. To my knowledge, no previous study has identified children with such low expected income ranks based solely on family background. Results for education and crime show that broader family information is even more valuable for measuring disparities in these outcomes.

²This analysis also shows that the rank-rank correlation in the Netherlands is higher than previously thought (Van Elk et al. (2024), Manduca et al. (2024), Boustan et al. (2025)). I provide a detailed explanation for why our estimates differ.

Using the Shapley value decomposition, I show how all family characteristics contribute to children's expected income. Nine of the ten most important variables relate to parental and extended family income and wealth, underscoring the central role of both economic resources and extended kinship in quantifying income gaps across family backgrounds. These variables matter for many children, but the decomposition also uncovers characteristics that affect smaller groups in profound ways. Parental absence, for example, contributes little to aggregate explanatory power but sharply reduces expected income for the children directly affected. The least advantaged children face cumulative disadvantages: their parents are often young, separated, and have low income and wealth, limited education, poor health, and criminal records, with similar disadvantages observed among aunts and uncles.

Lastly, I assess the roles of neighborhoods and pre-birth factors in driving the observed disparities. Neighborhoods play only a modest role: income gaps across family backgrounds are nearly as large within neighborhoods as between them, and neighborhoods explain little income variation beyond what is already explained by the observed family characteristics. Instead, a sizable share 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 a growing literature on the role of the broader family background in shaping children's economic success. Methodologically, I show how recent tools can be used to provide distributional insights even with many family background characteristics. Instead, prior studies that include multiple background variables have largely focused on summary statistics such as intergenerational mobility coefficients (Vosters and Nybom (2017), Vosters (2018), Adermon et al. (2021), Eshaghnia et al. (2022), Chang et al. (2025)), inequality of opportunity estimates (Brunori et al. (2023), Adermon et al. (2025)), or explanatory power measures (Mendolia and Siminski (2017), Blundell and Risa (2019)).³ I

³Brunori et al. (2024) also present distributional results using multiple family variables. However, their survey sample ($n = 7,297$) precludes analysis of the extremes of the population, and the family information is limited to parental education and occupation.

further demonstrate how the method of Lundberg et al. (2020) can be used to decompose income gaps across family background characteristics. Although this method is widely used in machine learning and is particularly suited for decomposing inequalities, it has, to my knowledge, not been applied in labor economics.

Empirically, I contribute by applying these methods to a large administrative dataset with the most detailed family background information studied to date.⁴ The large sample allows me to examine income disparities even among very small and (dis)advantaged groups of children, while the richness of the family information makes it possible to evaluate the relative importance of many dimensions that earlier work has considered only in isolation.⁵ This approach reveals strikingly large income gaps, especially for the most disadvantaged children. Since support for redistributive or opportunity-equalizing policies depend on the perceived role of family background in shaping economic outcomes (Alesina et al. (2018)), these findings provide important new evidence for the public debate.

Finally, this paper relates to work that measures the importance of the broader family background indirectly through income correlations among relatives, such as sibling correlations.⁶ While useful, such correlations mix family influences with unobserved factors like sibling spillovers, community influences, or common shocks, complicating interpretation. The prediction approach avoids this problem by relying only on observable family characteristics, whose contributions can be quantified. Moreover, sibling correlations only provide information on aggregate explanatory power but do not reveal the size of income gaps between specific family types.

This paper proceeds as follows. Sections 2 and 3 present the methodology and the data. Section 4 provides the main results and Section 5 explores drivers. Section 6 concludes.

⁴The closest is Blundell and Risa (2019), who use administrative data with parental income, education, net taxable wealth, occupation type, marital status, studying status, total work hours, household size, and region. In addition, I observe parents' specific assets and debts, health, crime, family structure, migration background, and also include aunts and uncles.

⁵This paper also stands out by including detailed information on mothers, rarely analyzed separately. Recent exceptions include Brandén et al. (Forthcoming), Ahrsjö et al. (2023), and Althoff et al. (2024).

⁶Other methods extend this approach to include more distant kinship (Collado et al. (2023)), information about parental income (Bingley and Cappellari (2019)), or schools and neighborhoods (Bingley et al. (2021)).

2 Methodology

This section illustrates how multiple background characteristics can be used to quantify income gaps across family backgrounds. I begin with a simple and widely used summary statistic, and then introduce two measures that provide more fine-grained insights. Finally, I show how recent advances in machine learning can be used to assess the relative importance of different family background variables for specific individuals.

Global measure. Let Y_{sf} be the income rank of a child s in a family f . Moreover, let $\mathbf{X}_f = (X_{f1}, \dots, X_{fK}) \subset \mathcal{X}$ be the set of all observable family background characteristics. Consider the following conditional expectation function decomposition of Y_{sf} :

$$Y_{sf} = E[Y_{sf} | \mathbf{X}_f] + \nu_{sf} = g(\mathbf{X}_f) + \nu_{sf}, \quad (1)$$

where, by construction, $E[\nu_{sf}] = E[\nu_{sf}m(\mathbf{X}_f)] = 0$ for any function $m : \mathcal{X} \rightarrow \mathbb{R}$.

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

$$R_{y|g}^2 = \frac{V(g(\mathbf{X}_f))}{V(Y_{sf})}. \quad (2)$$

I commonly refer to this metric as the ‘explanatory power’. Deutscher and Mazumder (2023) classify this as a global measure of intergenerational dependence, as it summarizes the importance of family background for the entire population.

The explanatory power is directly comparable to two commonly used alternative global measures of intergenerational dependence: the rank-rank correlation and sibling corre-

tions.⁷ I use this comparison to benchmark the explanatory power of the full model against (i) models based solely on parental income, which provide a lower bound, and (ii) models based on sibling fixed effects, which provide an upper bound.

Another closely related approach from the inequality of opportunity literature makes similar decompositions as in Equation 2, but typically uses other inequality measures than the variance. This is called the ex-ante approach to quantifying inequality of opportunity.⁸ This literature treats all background factors beyond an individual's control as 'circumstances'. The findings in this paper are specific to inequality of opportunity arising from family circumstances, a subset of all possible circumstances.

Local measures. I present two measures that provide more detailed insight into the size of the income gaps across family backgrounds. I first report the full distribution of expected incomes, $F(X) = P(g(\mathbf{X}_f) \leq X)$, which allows me to identify the expected income of the least and most advantaged children and all groups in between. This is the multidimensional analogue of commonly reported Conditional Expectation Function plots of child income given parental income. Next, I construct detailed matrices that map children with different expected incomes into quantiles of the observed income distribution. This is the multidimensional analogue of commonly reported transition matrices. In the framework of Deutscher and Mazumder (2020), these measures are local because they provide insights for specific subsets of children. Together, they provide the most granular evidence on how child income varies with a comprehensive set of family background characteristics.

Estimation. Estimation of the global and local measures requires estimation of the condi-

⁷This follows because in a regression of child income rank on parental income rank, there is a one-to-one relationship between the coefficient and its explanatory power ($R^2 = \beta^2$). Moreover, the sibling correlation equals the (adjusted) R^2 of a regression of child income on family fixed effects. Since these family fixed effects include all observed *and* unobserved factors shared between siblings, its explanatory power is weakly greater than that of a model which uses only observable factors shared between siblings (as in this paper).

⁸A 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.

tional expectation function $g(\mathbf{X}_f)$. A key challenge is that its functional form is unknown. Variables may enter in a non-linear manner or interact with other variables. In these cases, non-parametric machine learning methods outperform linear regression models. Accordingly, I employ gradient-boosted decision trees to generate these predictions (Friedman (2001)).⁹ Tree-based methods offer the additional advantage of providing Shapley value-based measures of variable importance even with a large number of predictors.

The complexity of tree-based models depends on several tuning parameters, such as the maximum number of splits per tree, the minimum gain required for a split, the total number of trees, and the learning rate. To select these parameters, I randomly split the sample into a training set (80 percent) and a test set (20 percent). I use 5-fold cross-validation on the training set to determine optimal values and then re-estimate the model on the full training set with these parameters. I apply the final model to the test set to obtain out-of-sample predictions, from which I compute both global and local measures.

Quantifying variable importance. Having identified children with low or high expected incomes, a next question is *why* children with specific covariate values have low or high expected incomes. Quantifying variables' contributions is challenging because for many methods it requires estimating a large number of models using all possible subsets of variables, which is an NP-hard problem. As a result, inequality decompositions are typically applied to small sets of variables and only decompose variables' contributions to aggregates like the R^2 or other inequality measures (Shorrocks (2013)).

Recent advances in machine learning make it possible to go further by applying novel Shapley value decompositions (Lundberg and Lee (2017)). Unlike traditional variance decompositions, this approach quantifies contributions to individual-level predictions. It can therefore reveal family background factors that are important for some individuals even if

⁹Single decision trees partition the covariate space into regions with similar outcomes and predict for new observations the average value in their region. Gradient-boosted trees improve on this by iteratively fitting trees to residuals, enabling them to capture non-linear relationships and complex interactions. Such tree-based methods have proven to be superior when predicting from tabular data (Grinsztajn et al. (2022)).

their aggregate contribution to R^2 is modest, a distinction that standard methods miss. This section discusses the method conceptually and Section 4.3 provides an application.

To fix ideas, define the marginal contribution of variable X_{fk} to a given subset of variables $S \subseteq \mathbf{X}_f \setminus \{X_{fk}\}$ by

$$h(X_{fk}, S) = E[Y_{sf}|S \cup X_{fk}] - E[Y_{sf}|S]$$

For example, when there are no other explanatory variables ($S = \emptyset$), then the contribution of X_{fk} equals $h(X_{fk}, \emptyset) = E[Y_{sf}|X_{fk}] - E[Y_{sf}]$. At the other extreme, when S is the set of all other variables, then the contribution of X_{fk} equals $h(X_{fk}, \mathbf{X}_f \setminus \{X_{fk}\}) = g(\mathbf{X}_f) - E[Y_{sf}|\mathbf{X}_f \setminus \{X_{fk}\}]$.

The Shapley value of X_{fk} for individual s in family f is the average of its marginal contributions, where the average is taken over all possible orderings of the covariates:

$$\phi_{sf}(X_{fk}) = \sum_{S \subseteq \mathbf{X}_f \setminus \{X_{fk}\}} w(S) h(X_{fk}, S),$$

where $w(S) = \frac{|S|!(K-|S|-1)!}{K!}$. For example, if parental income has a Shapley value of -2 for a given child, this means that including parental income lowers her expected income rank by two on average, where the average is taken over all possible subsets of included covariates.

Lundberg and Lee (2017) show that these Shapley values are the only measures of variable importance that preserve important properties from cooperative game theory.¹⁰ While exact Shapley values are computationally infeasible for most models due to the need to sum over all variable subsets, recent algorithms can compute exact Shapley values for tree-based models in short time periods (Lundberg et al. (2020)).¹¹ I use this algorithm to compute Shapley

¹⁰These properties are additivity and monotonicity. Additivity ensures that for a given input \mathbf{X}_f , the sum of the Shapley values equals the model's output $g(\mathbf{X}_f)$. Monotonicity guarantees that if a variable's contribution increases or stays the same, its Shapley value will not decrease, regardless of the other inputs.

¹¹Their key insight is that trees are particularly suited because moving down a path amounts to adding variables one by one through their splits. This structure makes it possible to track each variable's contribution to the prediction without re-estimating the model for all possible subsets.

values from the gradient-boosted decision tree, applied to a random sample of 10,000 children from the test set. Because the sample is randomly drawn, the resulting Shapley values are representative of the full population.

3 Data

Core analysis sample. I use administrative data from Statistics Netherlands covering the full Dutch population.¹² 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.¹³ Income is measured 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.¹⁴ 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

¹²Access is granted through a secure remote facility under a confidentiality agreement.

¹³Some 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.

¹⁴I exclude years with yearly household income below €1,000 (0.6%), as these cases typically correspond to wealthy entrepreneurs with business losses.

to age 34 for the youngest cohort (born in 1989). On average, children have nine income observations, with 96 percent having at least five. I then rank children within birth-years based on their lifetime household income. I also present results for various alternative measures to evaluate the sensitivity of the results due to these choices.

Children's education and crime. 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 C. 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.¹⁵

Second, despite the extensive coverage of variables, some missing values persist. Most importantly, education records for the parents' generation are incomplete. In a 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 Main Results

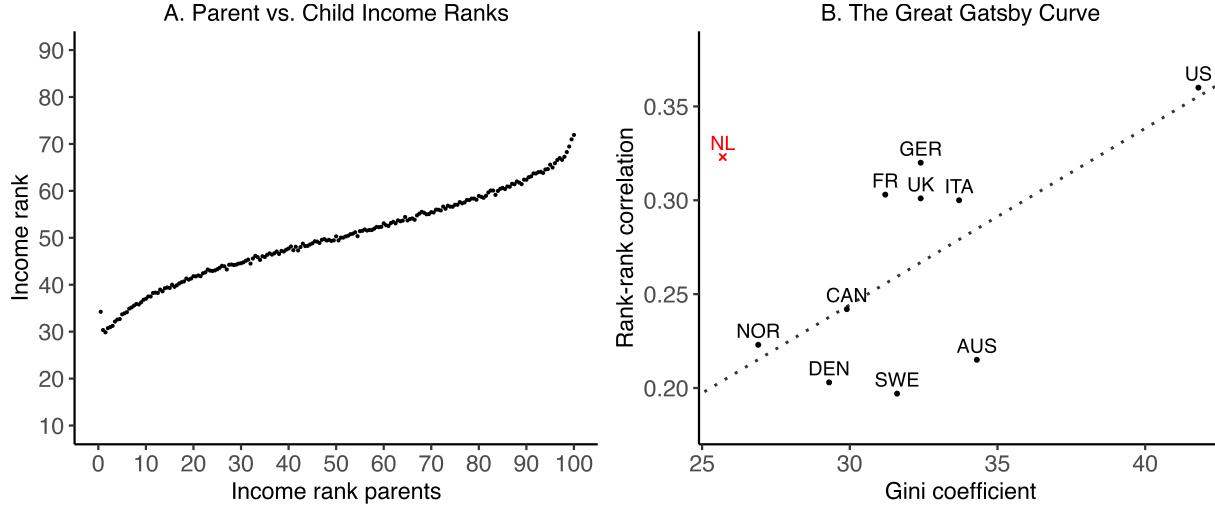
4.1 Intergenerational Income Mobility in the Netherlands

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

Figure 1 (a) 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

¹⁵This 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.

Figure 1: Intergenerational Income Mobility in the Netherlands



Notes: panel A presents a nonparametric scatter plot of mean income ranks versus parental income rank. The sample consists of all $N = 1,702,355$ children from the core analysis sample (Table A1) for whom parental income is not missing (99.1%). 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. Panel B presents a cross-country comparison of Gini coefficients and rank-rank correlations. The dotted line shows the regression line obtained when regressing rank-rank correlations on Gini coefficients, excluding the Netherlands. The Gini coefficients are taken from the most recent estimates between 2018 and 2023 from the World Bank. The rank-rank correlations are computed by: Heidrich (2017) (Sweden), Helsø (2021) (Denmark), Deutscher and Mazumder (2020) (Australia), Bratberg et al. (2017) (Norway), Corak (2020) (Canada), Kenedi and Sirugue (2023) (France), Dodin et al. (2024) (Germany), Acciari et al. (2022) (Italy), Rohenkohl (2023) (the UK), and Davis and Mazumder (2024) (the US). To make estimates comparable, the Italian estimate corresponds to the one when adjusting for lifecycle bias, incomplete coverage of taxpayers and tax evasion (reported on page 28). See Kenedi and Sirugue (2023) for a more detailed comparison of approaches.

income distribution.¹⁶ 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.¹⁷

While cross-country comparisons should be made with caution, Figure 1(b) suggests that the Netherlands ranks among OECD countries with relatively strong persistence. Its rank-rank correlation is higher than in Sweden, Denmark, Australia, Norway, and Canada (0.20–0.24), similar to France, Germany, Italy, and the UK (≈ 0.30), and below the United

¹⁶As 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.

¹⁷This estimate exceeds recent estimates for the Netherlands from Van Elk et al. (2024), Manduca et al. (2024), and Boustan et al. (2025), who report estimates between 0.16 and 0.23. In Appendix B, I replicate their approaches and illustrate why our estimates differ.

States (0.36). This is striking because the Netherlands has one of the lowest levels of income inequality, and lower inequality is often associated with weaker persistence. Indeed, Panel B suggests that the Netherlands is an outlier on the ‘Great Gatsby Curve’ (Corak (2013)).

Appendix B 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.

4.2 Including Detailed Parental Information

The previous section examined how child income varies with parental income. This section extends the analysis to a much broader set of family characteristics. I first assess the overall explanatory power of all family background variables and then provide two analyses offering more granular insights into income gaps across family backgrounds.

Explanatory power. 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 2. 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, mark-

ing a 58 percent increase compared to the income-only model (Figure 2).

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 .¹⁸ This 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, a source of measurement error that has received considerable attention in the literature (Mazumder (2005), Nybom and Stuhler (2017)).¹⁹ 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.

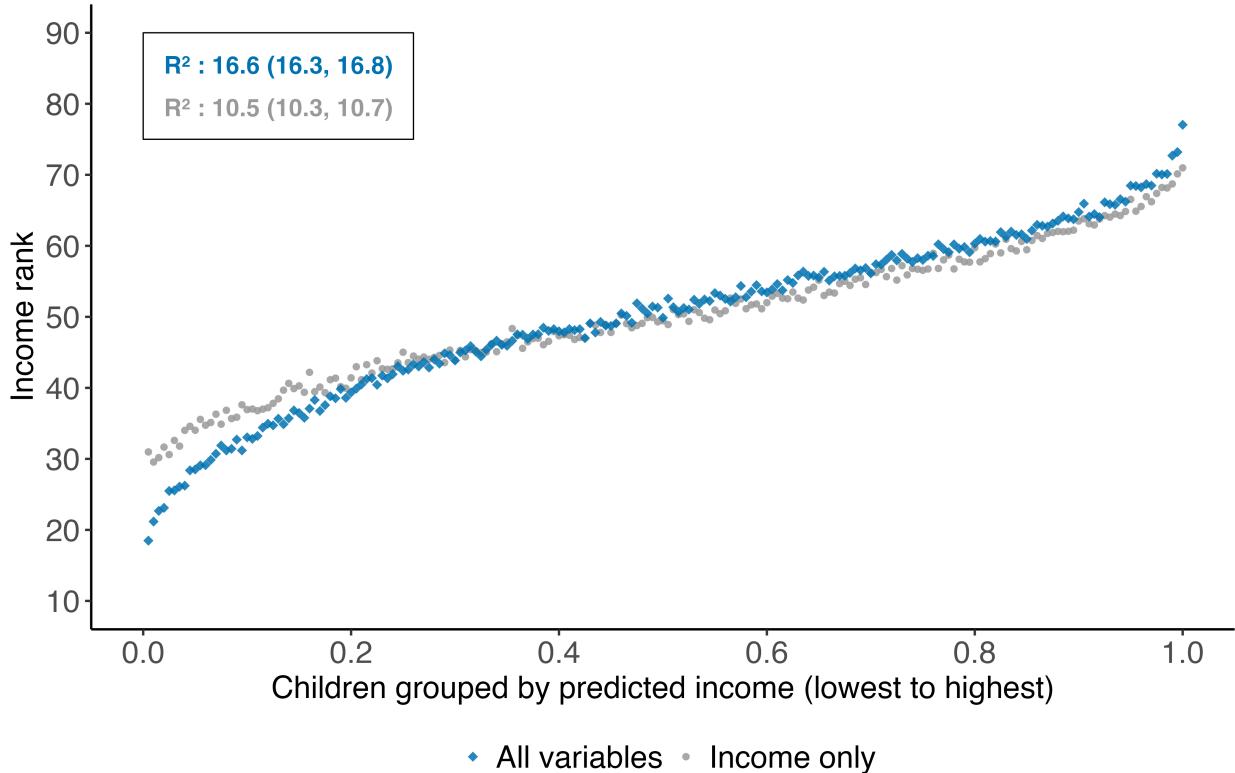
Distribution of expected income. Figure 2 provides a detailed binscatter plot of children’s mean income ranks, sorted from lowest to highest predicted income. The X-axis divides the test dataset into 200 bins, each containing approximately 1,700 children, based on their predicted income ranks within 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 grey diamonds reflect groupings based on predictions from the comprehensive model.

The comprehensive model identifies considerably greater income disparities by family background, particularly for the most disadvantaged children. 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.

¹⁸I use here that in a rank-rank regression, $R^2 = \beta^2$ (i.e. $0.408^2 - 0.324^2 = 0.166 - 0.105 = 0.061$).

¹⁹Table 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%.

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 blue 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 grey 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.

Transition probabilities. Figure 2 reports only children’s average income rank, yet there is substantial variation around this average. To capture this, Figure 3 (a) presents a 200×5 matrix with the shares of children in each adult income quintile given expected income. In the absence of any intergenerational transmission, the share of children in any income quintile should be 20 percent for all individuals. Instead, there are large differences. For instance, among the 0.5 percent children with the lowest expected incomes, less than two percent are in the top income quintile, whereas 69 percent are in the bottom income quintile.

To place these conditional probabilities in international perspective, panel B focuses on the 20 percent of children with the lowest expected income. I compare their probabilities of reaching the top quintile (‘moving up’) or being in the bottom quintile (‘staying down’) to estimates based on parental income alone, both for the Netherlands and for other countries. Even using only parental income, the Netherlands already shows strong persistence at the bottom: the twenty percent least advantaged children are nearly four times more likely to be in the bottom quintile than the top (35.8 vs. 9.3 percent).²⁰ Incorporating the full set of family background variables increases this ratio to almost seven (41.8 vs. 6.1 percent). These results indicate that for a substantial share of Dutch children, the chances of ‘moving up’ are strikingly low, and much lower than implied by parental income alone.

4.3 What Characterizes Family (Dis)Advantage?

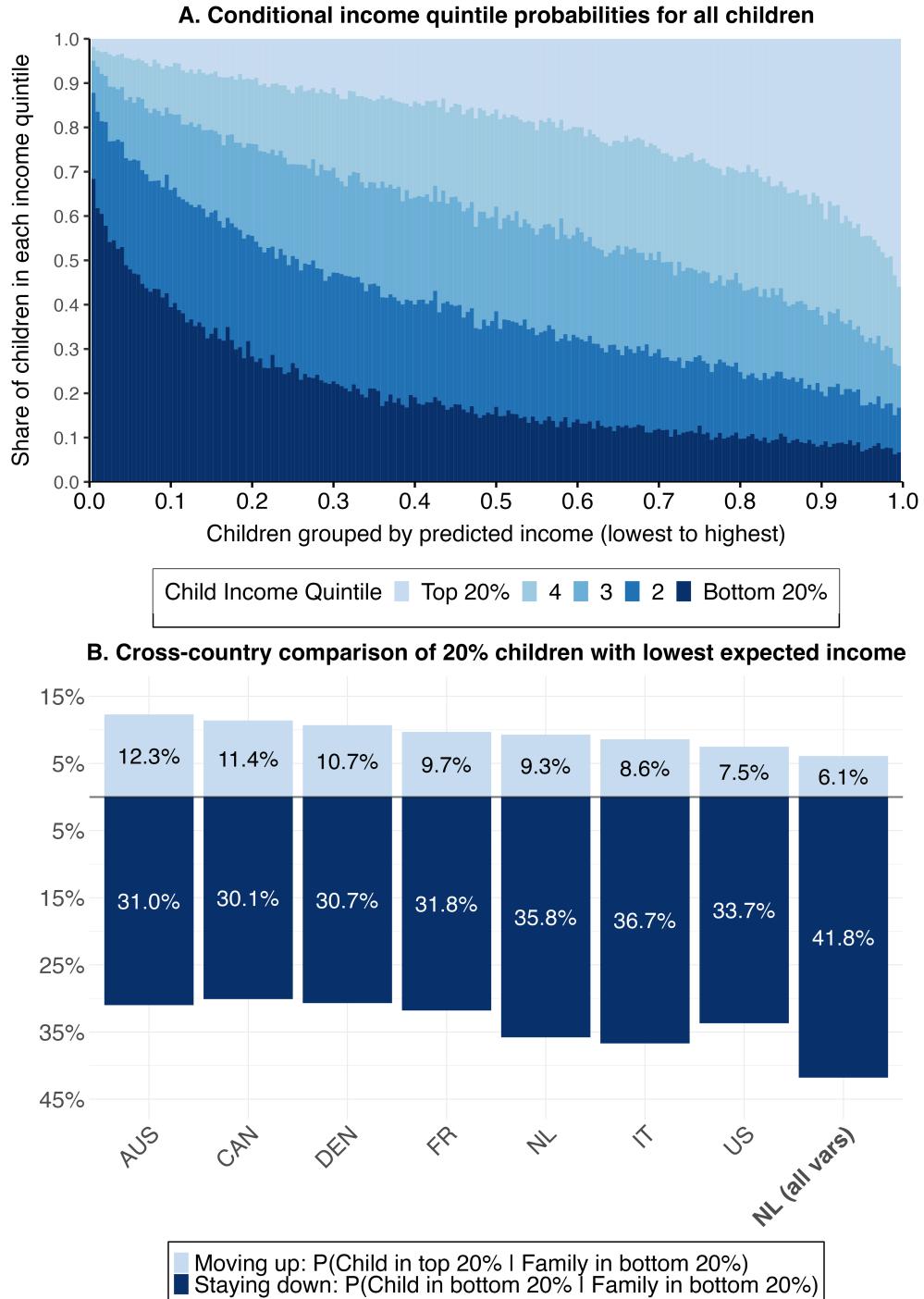
This section highlights which family characteristics strongly predict income and describes the characteristics of families at the extremes.

The best predictors. To understand which of the included family characteristics are most strongly associated with child income, I use the Shapley value decomposition described in section 2. Figure 4 presents a detailed graph illustrating the variable importance of the 30 most predictive variables, calculated using Shapley values.²¹ The boxplots report

²⁰In Figure B1, I provide the full 5×5 transition matrix based on parental income only.

²¹All variables contribute to the predictions, but the effects of the remaining variables are small and are

Figure 3: Child Income Distributions Across Family Backgrounds



Notes: Figure A shows the share of children in each income quintile. The sample contains all 340,608 children from the test data, and they are grouped into 200 equally sized bins according to their predicted income rank from the comprehensive model (as in Figure 2). Figure B shows the share of children in the top (transparent) or bottom (dark) income quintile. The first seven bars represent children from families in the bottom 20 percent of the parental income distribution. Estimates for other countries are taken from the studies listed below Figure 1. The final bar represents the 20 percent of children with the lowest predicted incomes from the comprehensive model (the bottom twenty percent in panel A).

the distribution of Shapley values for a randomly drawn subset of 10,000 children from the test data. As they are randomly drawn from the full population, these estimates are representative of the full population.

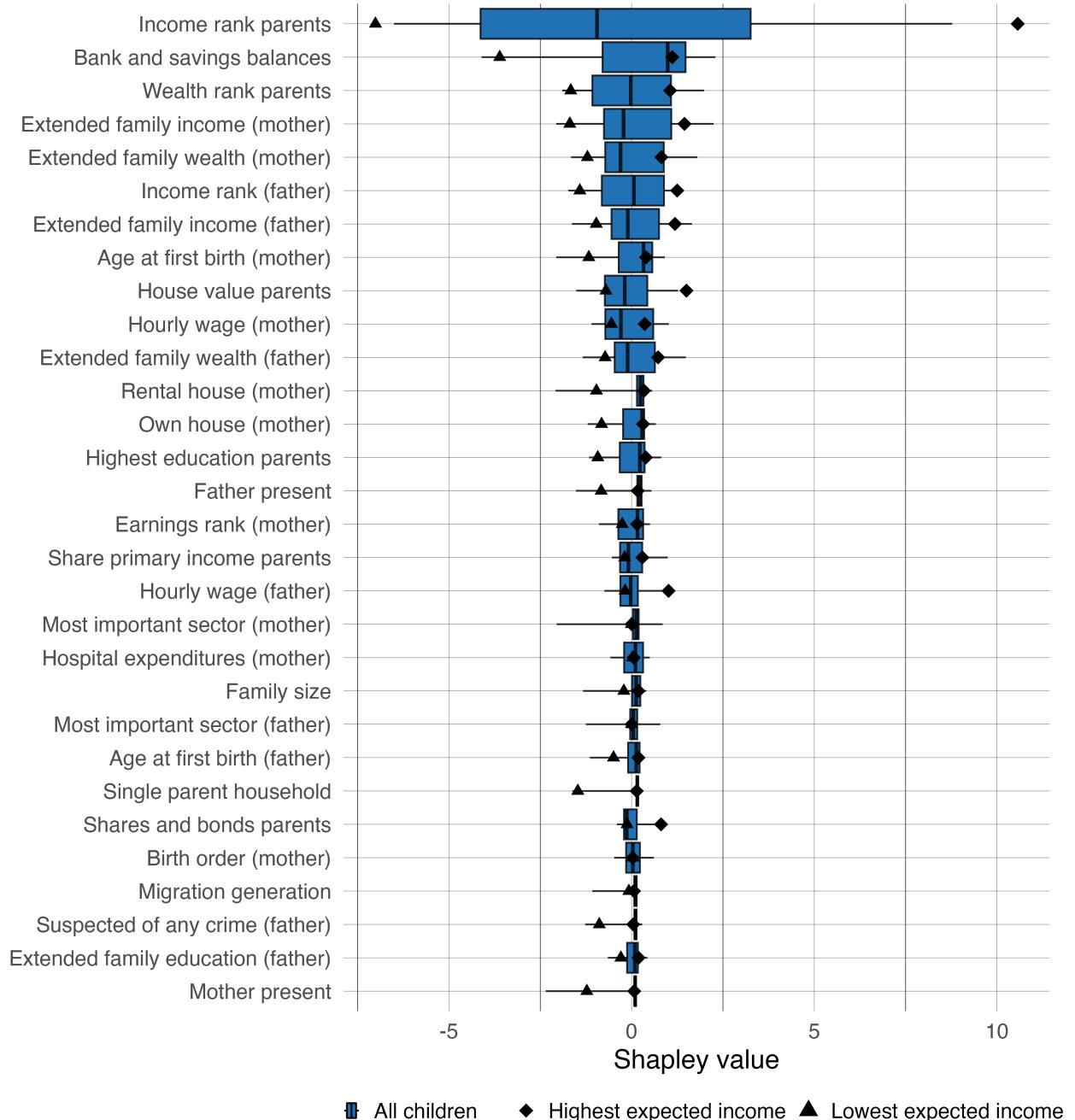
To illustrate their interpretation, consider the most important predictor of child income: parental income. The 2.5th percentile of the corresponding boxplot is -6.5, indicating that for 2.5 percent of the children, the Shapley value for parental income is below 6.5. As discussed in Section 2, a Shapley value of -6.5 means that including parental income to a set of other predictors on average reduces expected income by 6.5 ranks, where the average is taken over all possible subsets of predictors. The 75th percentile is 3.3, meaning that for 25 percent of the children, parental income increases expected income by more than 3.3 ranks.

Figure 4 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 expected income of many children. This underscores the importance of economic resources and extended family ties in capturing family-driven income inequality.

While other variables contribute less on average, they can make sizeable contributions for small subsets of children. To illustrate this, consider the Shapley values for mother presence. The boxplot is centered around zero, indicating small contributions to expected income for most children. This is expected, since 96 percent of children live with their mother at age 15. However, the left whisker shows that for 2.5 percent of children, whose mothers are absent, expected income decreases by at least 2.4 ranks. Thus, although mother presence contributes little in the aggregate, its absence has a substantial negative effect on a small group. Such effects would be overlooked when focusing only on aggregate explanatory power, as in traditional variance decompositions.

Children at the extremes face a combination of (dis)advantages that accumulate into their exceptionally high or low expected incomes. The triangles report mean Shapley values for therefore omitted for conciseness.

Figure 4: The 30 Most Predictive Family Characteristics of Child Income



Notes: each row displays a boxplot representing the distribution of Shapley values for a given variable. These 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. 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. The triangles (diamonds) report the average Shapley value for each of those variables for the 0.5 percent of children ($n = 1703$) from the test data with the lowest (highest) predicted incomes.

the 0.5 percent of children with the lowest predicted incomes. The average Shapley value for parental income is -7, while the sum of all Shapley values must be close to -32 to match their average income rank of 18. This implies that their low expected income is driven by multiple family characteristics that jointly push it downward. The figure illustrates this pattern: the average Shapley values are negative for many variables, including some that contribute little in the aggregate, such as living in a single parent household or having a father suspected of a crime. All these variables jointly result in these children's low expected income.

Descriptive statistics. To further illustrate these cumulative (dis)advantages, Table 2 provides descriptives of children with different expected incomes.²² 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 family characteristics of the children at the bottom of the expected income distribution are unfavorable in nearly every measurable dimension. 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 multiple family background characteristics.

4.4 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

²²As 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.

Table 2: Family Background Characteristics across the Predicted Income Distribution

	<i>Predicted Income Bins</i>								
	0-	0.5-	1-	5-	10-	90-	95-	99-	99.5-
	0.5	1	5	10	90	95	99	99.5	100
Child income rank	18.28	21.28	25.78	31.04	50.53	65.82	69.58	73.29	77.58
<i>Family background characteristics</i>									
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.

the generalizability of the main results.

As violent crimes are predominantly committed by men (85 percent), I focus on men's criminal behavior only. See Figure A1 for women's criminal behavior. Moreover, as discussed in Section 3, both analyses are restricted to children born between 1985 and 1989.

Figure 5 (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.²³ This increase in explanatory power considerably exceeds the 58 percent improvement observed when predicting child income (Figure 2).

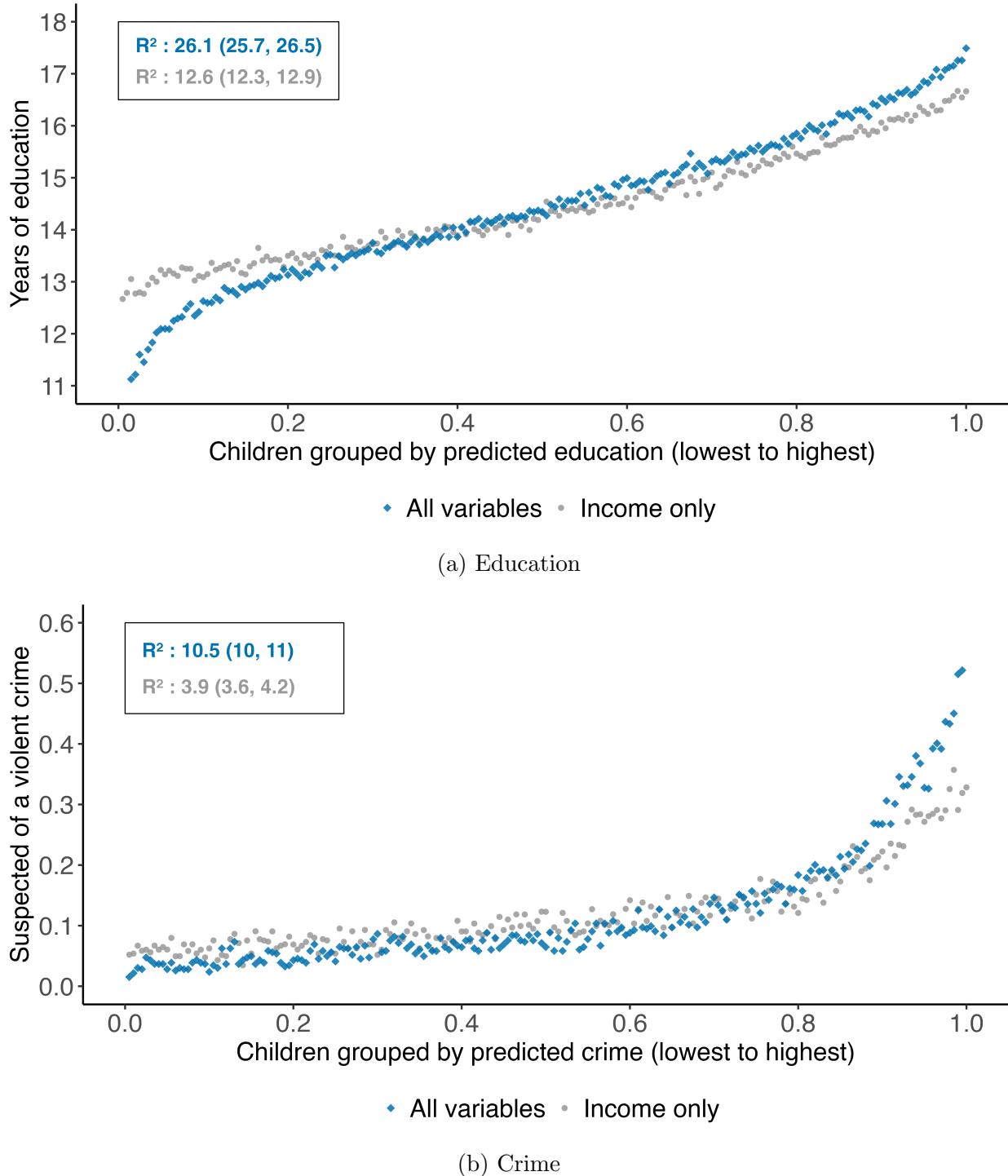
This is also true for predicting boys' violent crime. Figure 5(b) shows that the explanatory power for predicting crime increase from 3.9 percent for the model that incorporates income only to 10.5 percent for the comprehensive model, marking a 169 percent increase. Taken together, it appears that including 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 5 (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.²⁴

²³Intergenerational 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.

²⁴Other works have mostly focused on intergenerational crime associations (see Besemer et al. (2017) for a review) or sibling correlations Eriksson et al. (2016).

Figure 5: Predicting Children's Education and 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 blue and grey 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.

4.5 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 0.308 (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 (0.166/0.308). The remaining half of siblings' similarities 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.²⁵ 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 somewhat diminishes the impact of family (dis)advantages.

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

²⁵Disposable income is the amount left after deducting taxes and social insurance payments from gross income.

achieves an explanatory power of 15.3 percent.²⁶ This is quite 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. However, a downside of OLS regression is that it does not support the computation of individual-specific Shapley values.

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.²⁷ 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.

²⁶Coefficient estimates are available upon request.

²⁷In 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.

5 Drivers of Family (Dis)Advantage

In this final section, I use the predictions from the comprehensive model to explore the importance of two channels. First, I study the role of neighborhoods. Second, I decompose the observed disparities into shares driven by post-birth factors versus pre-birth factors.

5.1 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 4 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 15.²⁸ 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.²⁹ 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

²⁸The 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.

²⁹I use that the variance in predicted incomes is $V(\hat{Y}) = 125$ and that the neighborhood fixed effects explain 30 percent of this variance: $V(\alpha_N)/V(\hat{Y}) = 0.3$. Then it follows that $SD(\alpha_N) = \sqrt{0.3 \cdot 125} = 6.1$.

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.³⁰

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.³¹

Table 3: Predicted and observed income differences between neighborhoods

	Predicted income (1)	Income (2)	Income (3)
Predicted income	-	1.003 (0.004)	0.943 (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 neighborhood 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.

5.2 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,

³⁰These 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)).

³¹Note that this exercise is conceptually distinct from mover designs such as Chetty and Hendren (2018) and Deutscher (2020). Those studies examine whether regional *differences* in persistence are causal, rather than estimating the share of *overall* persistence attributable to neighborhoods.

children born into advantaged families may benefit from favorable genetic endowments for educational attainment or higher quality prenatal care, both of which can shape long-term income trajectories. To get insight into their relative importance, this section decomposes the observed disparities into shares due to pre-birth factors versus post-birth factors.

To do this, 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.³² 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.³³

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

³²Although 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.

³³Sacerdote (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.

waiting times.³⁴ 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 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 family (dis)advantage.

³⁴Waiting times during this period could span several years. See, for example, the government report ‘Rapport Commissie Onderzoek Interlandelijke Adoptie’ (in Dutch, 2021).

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 (0.028)	0.282 (0.029)	0.285 (0.031)	0.282 (0.027)	0.872 (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	Own-birth children in 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.

6 Conclusion

This paper measures income gaps across family backgrounds. I show how recent machine learning tools can provide detailed evidence on income distributions across multiple family background characteristics and quantify their importance at the individual level. I apply these methods to rich Dutch administrative data containing the most comprehensive family information studied to date. The results show that standard intergenerational mobility measures substantially underestimate the (dis)advantages children face, especially among the most disadvantaged. Their chances of upward mobility are far lower than analyses based solely on parental income suggest. To capture these disparities more accurately, detailed data on parental and extended family income and wealth prove particularly valuable.

This multidimensional approach is best viewed as a complement to 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 underestimate 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 expe-

rienced by different children and what characterizes their families, this paper demonstrates the value of a richer, multidimensional approach. As more comprehensive data becomes accessible, future research can employ this approach to further explore the intricate intergenerational transmission process.

References

- Acciari, Paolo, Alberto Polo, and Giovanni L. Violante.** 2022. “And Yet It Moves: Intergenerational Mobility in Italy.” *American Economic Journal: Applied Economics* 14 (3): 118–163.
- Adermon, Adrian, Gunnar Brandén, and Martin Nybom.** 2025. “The Relationship between Intergenerational Mobility and Equality of Opportunity.” IFAU Working Paper No. 2025:2.
- Adermon, Adrian, Mikael Lindahl, and Mårten Palme.** 2021. “Dynastic Human Capital, Inequality, and Intergenerational Mobility.” *American Economic Review* 111 (5): 1523–1548.
- Ahrsjö, Ulrika, René Karadakic, and Joachim Kahr Rasmussen.** 2023. “Intergenerational Mobility Trends and the Changing Role of Female Labor.” arXiv preprint, arXiv:2302.14440.
- Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso.** 2018. “Intergenerational Mobility and Preferences for Redistribution.” *American Economic Review* 108 (2): 521–554.
- Althoff, Lukas, Harriet Brookes Gray, and Hugo Reichardt.** 2024. “The Missing Link(s): Women and Intergenerational Mobility.” Stanford University Working Paper, https://lukasalthoff.github.io/pdf/igm_mothers.pdf.
- Besemer, Sytske, Shaikh I. Ahmad, Stephen P. Hinshaw, and David P. Farrington.** 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, 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. “Inter-

- generational 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, Guido Neidhöfer, and UNU-WIDER.** 2023. “Inequality of Opportunity and Intergenerational Persistence in Latin America.” WIDER Working Paper 2023.
- Brunori, Paolo, Francisco H.G. Ferreira, and Pedro Salas-Rojo.** 2024. “Inherited Inequality: A General Framework and a ‘Beyond-Averages’ Application to South Africa.” IZA Discussion Paper No. 17203.
- 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, and Nathaniel Hendren.** 2018. “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects.” *The Quarterly Journal of Economics* 133 (3): 1107–1162.
- 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 Ortúñoz-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.** 2013. “Income Inequality, Equality of Opportunity, and Intergenerational Mobility.” *Journal of Economic Perspectives* 27 (3): 79–102.
- 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.** 2020. “Place, Peers, and the Teenage Years: Long-Run Neighborhood Effects in Australia.” *American Economic Journal: Applied Economics* 12 (2): 220–249.
- 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.
- Dodin, Majed, Sebastian Findeisen, Lukas Henkel, Dominik Sachs, and Paul Schüle.** 2024. “Social Mobility in Germany.” *Journal of Public Economics* 232 105074.
- 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.
- Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux.** 2022. “Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?.” July.
- 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.
- Mendolia, Silvia, and Peter Siminski.** 2017. “Is Education the Mechanism through Which Family Background Affects Economic Outcomes? A Generalised Approach to Mediation Analysis.” *Economics of Education Review* 59 1–12.
- 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 Neighbour-

- hood 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.
- Shorrocks, Anthony F.** 2013. “Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value.” *The Journal of Economic Inequality* 11 (1): 99–126.
- 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
<i>Most important source of income</i>					
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
<i>Type of housing</i>					
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
<i>Extended family outcomes</i>					
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.³⁵ 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.

³⁵One 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 C1. 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 (1)	Test data sample size (2)	R^2 (3)	0.025% lower bound (4)	97.5% upper bound (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

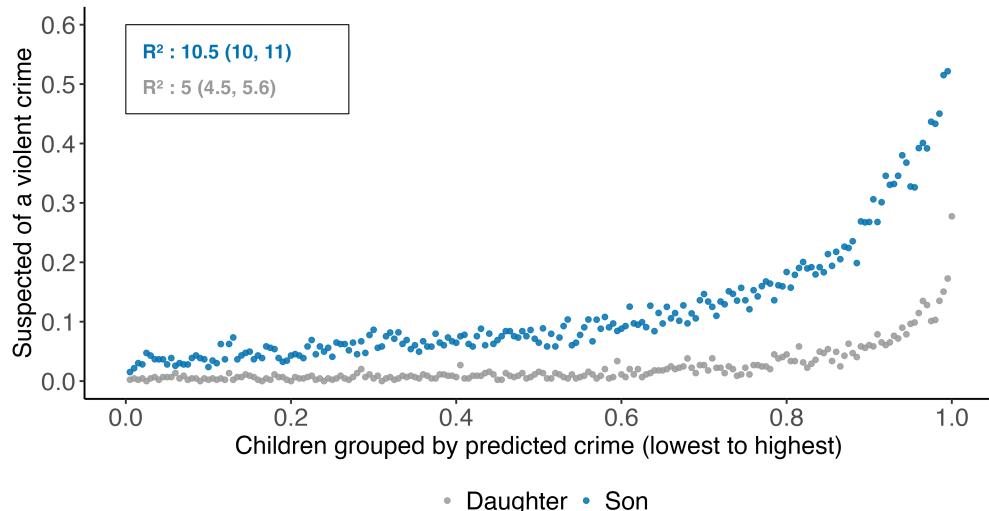
Notes: This table presents estimates of explanatory power for gradient-boosted decision trees that include all explanatory variables (as in Figure 2), using smaller samples. Column 1 reports the share of the core sample that is used for the analysis. Column 2 reports the sample size of the test-data.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

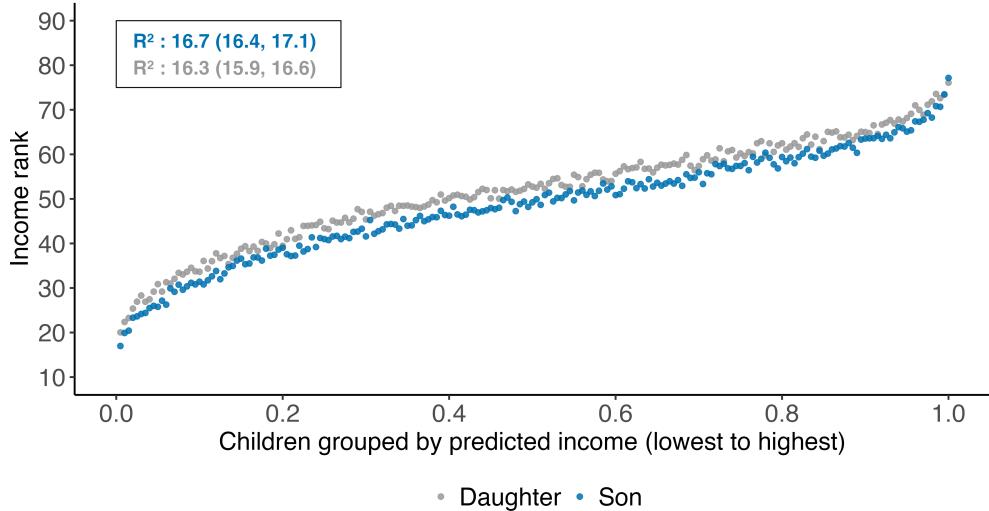
Notes: Each row presents the R^2 and corresponding 95% lower and upper bound for gradient-boosted decision trees that include all explanatory variables to predict child income (as in Section 4). The analysis sample consists of all 330,018 children born in 1980 and 1981 for whom I observe all incomes between ages 30 and 41. Each model is trained on the same randomly selected 80% of this sample, and evaluated on the remaining 20%. Panel A varies the number of years of income data used to construct the child income rank. The 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



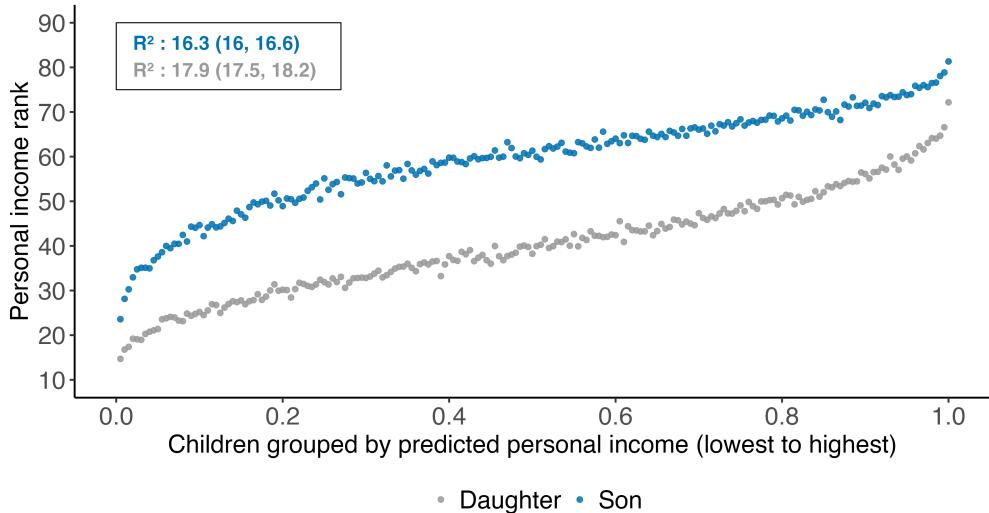
Notes: this figure presents a scatter plot of predicted crime for 92,725 sons and 89,051 daughters separately. Crime is measured as an indicator that equals 1 if a child was suspected of a violent crime between the ages of 20 to 33. The graphs are constructed using the same steps as in Figure 2. Confidence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



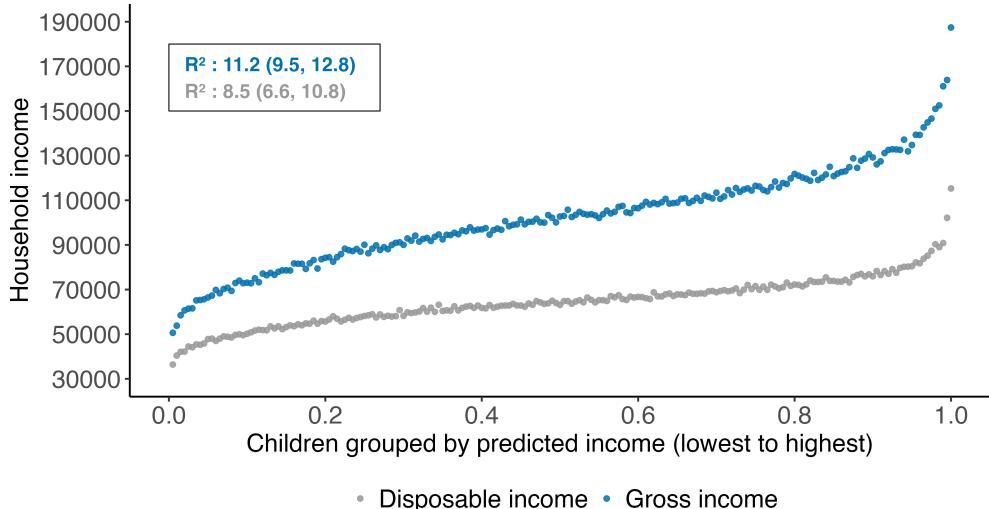
Notes: this figure presents binscatter plots of sons' and daughters' household income ranks for 173,652 sons and 166,957 daughters in the test data, who are sorted into bins based on their predicted income rank. Predictions are generated using the same predictive model and explanatory variables as in Section 4, 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
<i>Characteristics Adoptive Parents</i>										
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, which equals the adjusted R^2 of a regression of child income on sibling fixed effects. This estimate suggests that about 31% of all variation in income ranks is driven by factors shared between siblings.

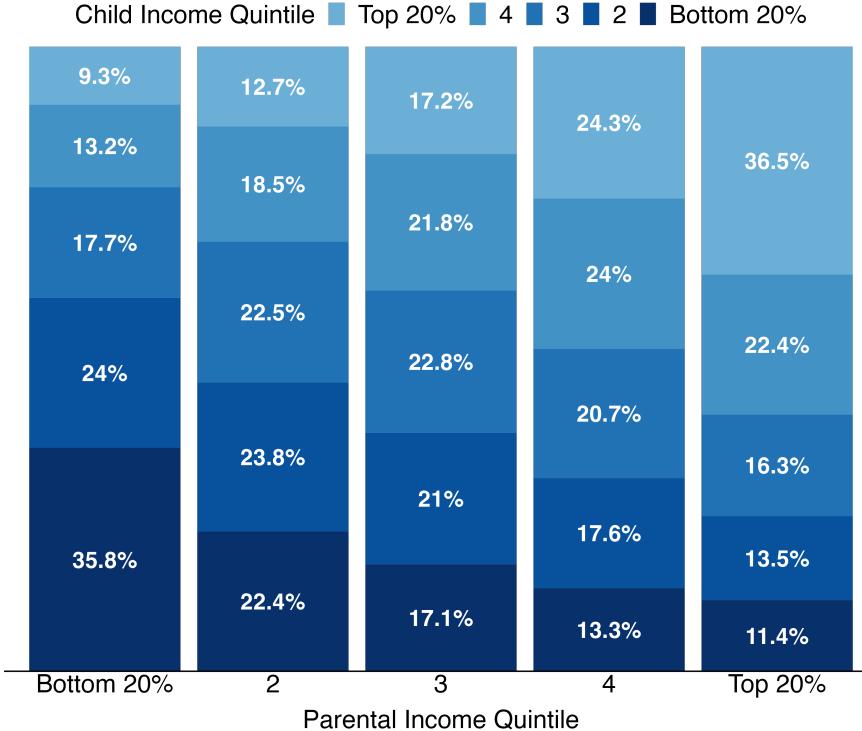
Figure B1 reports a 5×5 transition matrix. This table can be used to compare upward or downward mobility estimates across countries.

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 (0.001)	0.324 (0.001)	0.288 (0.001)	0.290 (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. 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 ($N = 1,702,355$) 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 (0.001)	0.304 (0.001)	0.311 (0.001)	0.316 (0.001)	0.320 (0.001)	0.323 (0.001)	0.325 (0.001)	0.327 (0.001)	0.329 (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 (0.001)	0.294 (0.001)	0.292 (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 (0.002)	0.304 (0.002)	0.308 (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, as documented in the 'parent-child register'. 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.³⁶

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

	(1)	(2)	(3)	(4)	(5)	6)
Coefficient	0.329 (0.001)	0.313 (0.001)	0.306 (0.001)	0.285 (0.001)	0.256 (0.001)	0.201 (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 not 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.

³⁶Using 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.³⁷ 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.

³⁷Estimates 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 (0.001)	0.292 (0.001)	0.256 (0.001)	0.229 (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: 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 C1. 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.

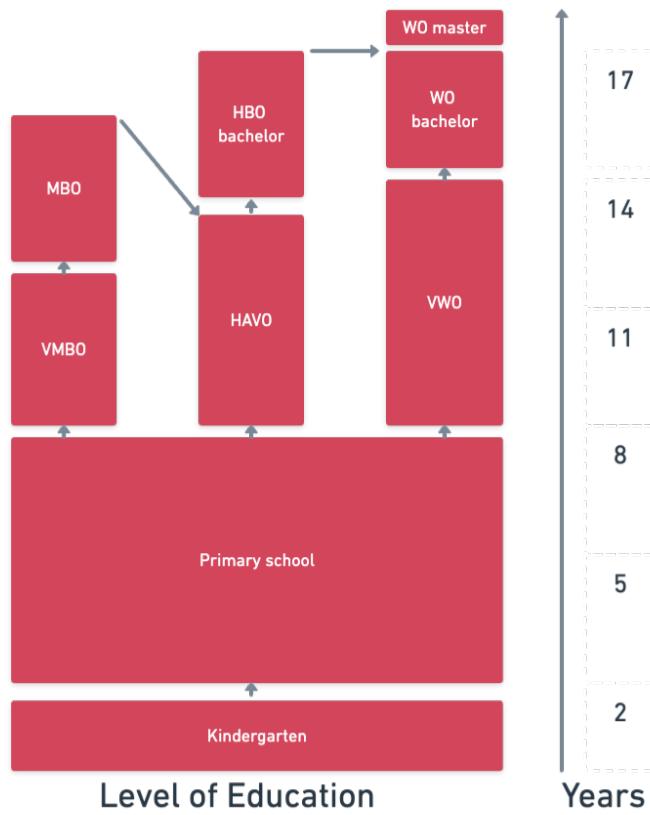


Figure D1: The Dutch Educational System

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 C1: 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