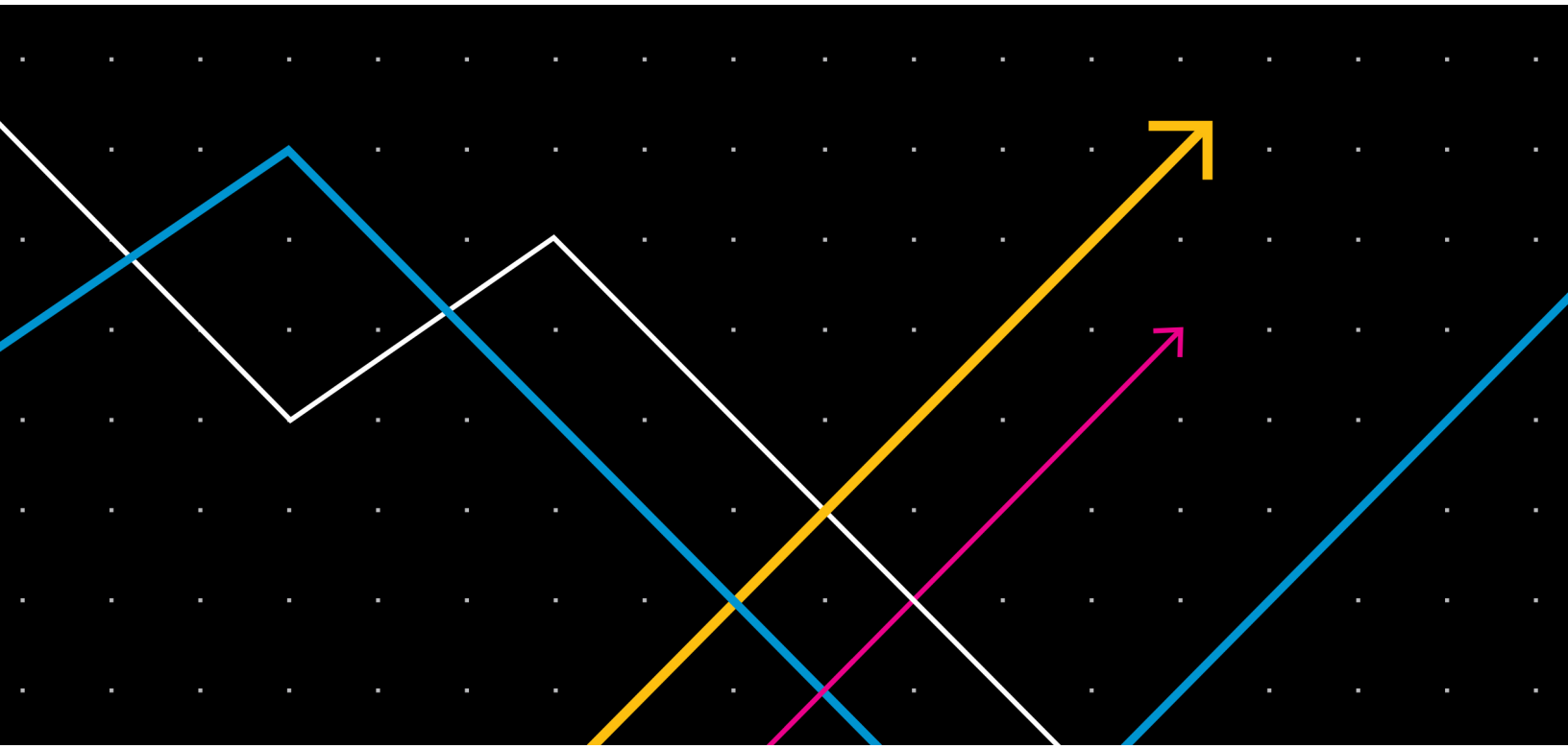

TECHNICAL DOCUMENT

# Social Genome Model 2.1

## Technical Documentation and User's Guide

*Kevin Werner*
URBAN INSTITUTE

*Kristin Blagg*
URBAN INSTITUTE

*with Gregory Acs and Steven Martin*
URBAN INSTITUTE

*and Alison McClay, Kristin Anderson Moore, Gabriel Piña, and Vanessa Sacks*
CHILD TRENDS

*November 2022*

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Child TRENDS


## ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

# Contents

# Acknowledgments

The Social Genome Model, originally developed by Isabel Sawhill at the Brookings Institution, is now a partnership between the Brookings Institution, Child Trends, and the Urban Institute. The current version of the model (2022) was developed by Child Trends and Urban under grants from the Chan Zuckerberg Initiative and the Bill and Melinda Gates Foundation.

The views expressed are those of the author and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors thank the Social Genome Project's advisors (Jessica Banthin, Jennifer Brooks, Kenneth Dodge, Greg Duncan, Kathleen Romig, John Sabelhaus, and Isabel Sawhill) for the input and advice as we built the current (2022) version of the Social Genome Model. We also thank Cary Lou for conducting rigorous code checks, and Kyle Ueyama, Michelle Menezes, and David D'orio for developing a user interface for the model. In addition, Kristin Moore, Vanessa Sacks, Gabriel Piña, Alison McClay, and Jon Schwabish provided thoughtful comments on earlier drafts.

# New Edition Note

This technical document and user's guide is an update to the previous edition, *Social Genome Model 2.0 Technical Documentation and User's Guide.* This version describes the new model and includes sample results and summary statistics that reflect it; this version incorporates new variables and removes old ones. This version supersedes and replaces the 2.0 edition.

# Executive Summary

**The Social Genome Model (SGM) is a lifecycle model that uses data from three longitudinal surveys to track a matched panel of individuals from birth to age 30. The goal is to understand how private and public policy interventions could improve lifetime outcomes of children and young adults. The model also allows researchers to track patterns of development across different gender and racial or ethnic groups.**

This technical document outlines the process of creating the Social Genome Model, version 2.1. First, we give an overview of the conceptual framework behind the model, then describe the two main datasets used: the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K), and the National Longitudinal Survey of Youth 1997 (NLSY). Next, we explain the process of matching observations across the ECLS-K and NLSY datasets to create the matched panel and the validation for our matching approach. We then show summary statistics for the variables included in our final dataset and discuss the parameterization of the model. We also discuss how we make use of estimates from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) to link early childhood (age 2) to the preschool years (age 5).

# Conceptual Framework

At different points in an individual's life—what we refer to as "life stages"—we measure key developmental outcomes and the factors and life contexts that influence those outcomes. This "ecological" model of development is widely accepted by practitioners and developmental researchers; it posits that development is a function of a variety of influences (Bronfenbrenner 1979). These influences include the individual's own characteristics; the characteristics of the family and household; child care or educational settings, peers, and neighborhoods; and the larger social context. Recent research documents that neighborhood characteristics such as concentrated poverty and crime rates influence children's adult outcomes (Chetty and Hendren 2018).

The "life course" model in turn posits that outcomes at any given life stage are influenced by factors from earlier life stages (Elder 1998; Shonkoff and Phillips 2000). For example, being born into poverty will potentially influence a child's cognitive development and other outcomes in early childhood, as well as outcomes through all subsequent life stages, culminating in how economically successful that child is at age 30.

The model is predicated on a "whole child" perspective that identifies multiple developmental domains at each stage of life, including cognitive, social, emotional/mental, health, and relationships (Moore 1997, 2020; Moore et al. 2017). It is also informed by human capital theory and literature documenting the importance of both cognitive and noncognitive skills in early and later childhood for achieving widely held measures of success in adulthood (Duckworth and Seligman 2005; Heckman and Rubinstein 2001; Heckman, Stixrud, and Urzua 2006; Shonkoff and Phillips 2000). The model can simulate how changes and interventions during an earlier developmental stage may ripple through a child's life.

Circumstances at birth: parents matter. Parents determine a child's home environment and genetic endowment. An extensive literature uses multiple measures of "class" or assessments of advantages and disadvantages at birth. Of these measures, maternal education approximates some mixture of genetic endowment and home environment. Parents' rank in the income distribution is one way to look at family background. The child's birth weight, as a proxy for prenatal environment, can be critical to future development (Glover 2011). In addition, health conditions can affect children's prospects. The mother's age at her first birth and family structure are also important: children of older and continuously married parents have more favorable mobility patterns than other children, partly

owing to higher incomes and more engaged parenting but also because of other advantages correlated with marriage (Hoffman and Maynard 2008; Sawhill 2014).

- Childhood and adolescence: development continues at school.Although human development begins in the home and is greatly influenced by parenting, the process continues in preschool and school (Garcia and Heckman 2014). We measure the acquisition of a broad set of skills throughout schooling years. We look at students' math and reading ability as well as grade point averages (GPAs) and educational attainment (graduation from high school or college). We also include indicators of internalizing and externalizing behavior, health, interpersonal relationships, school suspension, and involvement in crime, all factors that can be measured and directly impact subsequent life success.[1]

Into adulthood: income and adult success. The definition of success can instigate deep normative questions, with respect to defining what it means to successfully transition to adulthood or become "middle class by middle age." Income is a common measure of such success, although it is not a straightforward. For example, there are issues about whether to focus on the individual or the family and whether or how to adjust for family size. Some scholars prefer to define success using a measure of capacities (such as health and education) over income (Sen 1992). Ross, Moore, and colleagues have identified a measure of job quality that goes beyond income to also include fringe benefits, reasonable hours, and personal satisfaction (Ross et al. 2018). The SGM includes measures of both adulthood income and as well as health. The SGM also includes a measure of lifetime earnings, assessed for age 65, generated using data from the Urban Institute's DYNASIM model.[2] Lifetime earnings is based on education, health, and earnings at age 30. Including lifetime earnings allows us to see how changes in circumstances can affect this one measure of success over the course of an individual's entire life, even after the model ends.

# Success across Developmental Stages

The SGM is structured as a series of regression equations in which outcomes at each life stage potentially depend on the outcomes at all prior life stages. Having defined stages in the life course for birth, childhood/schooling, and adulthood, the SGM allows for interventions at each stage. A body of literature describes ways in which interventions earlier in life are related to success later in life.

Success begets success. To paraphrase economist James Heckman (2000), success begets success. That is, human capital formation is cumulative, and rates of return vary with prior skill development. In

other words, to succeed at a given life stage, it is helpful to have succeeded at the previous life stages. Also, varied types of skills are often complementary. A classic demonstration of this principle was the success of the HighScope Perry Preschool Project, a collaborative program started in 1962 of high-quality preschool education and weekly home visits for Black children in families with poverty-level incomes. The program was evaluated in a randomized controlled trial in which some children received the educational program and others did not. The children in the Perry Preschool Project acquired noncognitive skills that helped them focus on developing cognitive skills and as a result, did better through high school and into adulthood as measured by educational attainment and economic outcomes (Heckman et al. 2010). This finding is one reason Cunha and Heckman (2009) conclude that later-stage interventions designed to remediate early-stage deficiencies are costlier than earlier interventions.

Early success can languish. The full benefits of early-stage interventions will often not materialize without some investment during later stages. Currie and Thomas (1995) show that participants in the Head Start program lose some of their performance advantage over nonparticipants after aging out of the program. The Chicago Longitudinal Study, which tracked children in a preschool program, also found that adolescent and adult-stage benefits were greater for children who received extended interventions through sixth grade; later investment helped the children capitalize on earlier investment (Reynolds et al. 2011). As described in later sections, an advantage of the SGM is that it can capture effects of sustained interventions across childhood and adolescence.

Early interventions can have benefits that reemerge later in life. The effects of interventions at an early life stage might leap over life stages, affecting outcomes at a later life stage net of any effects that can be measured at the adjacent life stage. Most research on this topic has focused on negative "sleeper effects" such as neurological or other damage that has no immediate effect but can disrupt educational success later in life (Nelson and Magnuson 2011). The SGM can capture many but not all of the pathways by which early interventions might have lasting direct effects two or more life stages later.

Each of the three processes described above will produce a distinct pattern that the SGM is designed to capture: a process of "success begets success" will produce a one-step-at-a-time pattern whereby a variable in one life stage has a strong relationship to an outcome in the next life stage, which in turn has a strong relationship to an outcome in the following life stage, and so on to the final outcome of interest. A process of early "success languishing" will start as above, but at some later life stage the chain of relationships will break so that the initial intervention has little or no association with the final outcome. In a process involving a reemergence of effects, a variable at an early life stage

is connected to the final outcome not only by a series of adjacent steps, but also by a direct relationship to that later life stages. Thus, a covariate in an early life stage can produce a strong impact for an outcome two or more life stages later.

## Relationship between Education and Earnings

Research on human capital development provides insights on how interventions at one stage of a person's life can influence outcomes later in life. Although our intent is to measure human capital broadly to include health, attitudes, and habits, at the core of the model is the relationship between education and earnings in the tradition of Becker (1975), Mincer (1981), and later contributors to the human capital literature. The SGM's form and use are informed by lessons from that literature on the earnings returns to education:

**Education is important**. The rate of return on a year of schooling is generally found to be about 6 to 10 percent (Patrinos 2016). Recent research found that rates of return from education have increased for current cohorts compared to earlier ones, possibly because of a lag in the response of supply to demand (Goldin and Katz 2008).

**Returns vary**. Marginal returns may differ from average returns and depend on who is being targeted by an intervention (Carneiro, Heckman, and Vytlacil 2011). Rates of return vary by subgroup, with Black people experiencing higher returns than white people, and youth experiencing higher returns than the elderly (Henderson, Polachek, and Wang 2011). The rate of return to education is also heterogeneous across skill sets and depends on labor market demand (Sawhill and Owen 2013).

**Estimates are often reasonable.** Most of the results from ordinary least squares regressions reflect a causal effect, not ability bias; that is to say, higher earnings are the result of additional education and not reflective of underlying, innate ability that contributes to both higher educational attainment and higher earnings. The ability bias in such estimates is small and likely compensated by a bias in the opposite direction caused by measurement error (Card 2001).

**Much remains unexplained.** Individual and family earnings are an important element to an individual's success, and in inputs to success later in life. Yet individual earnings do not only depend on human capital accumulation broadly defined but are also determined by imperfections in the labor market (e.g., discrimination or high rates of unemployment induced by a recession) and each individual's unobserved characteristics. Therefore, even well-specified earnings equations explain only a modest portion of the variance in individual earnings.

These findings on the relationship between education and earnings do not prove similar relationships exist across variables within the SGM. However, they provide some guidance for interpretation and a basic proof of concept for the approach. Below, we describe how we developed the unique dataset used to capture key information on children and young adults from birth through age 30. We then detail how we estimated the regression equations that underlie the model.

# Dataset Cleaning and Assembly

The SGM uses data primarily from two sources: the Early Childhood Longitudinal Study, Kindergarten Class of 1998–99 (ECLS-K) and the National Longitudinal Survey of Youth 1997 (NLSY97). These datasets are described in more detail below. The model also uses the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), but unlike with the ECLS-K and NLSY97, data from the ECLS-B survey are not used in the SGM data set directly. Instead, the model includes coefficients estimated from *outside the model* using the ECLS-B to describe relationships between early childhood and preschool variables. The ECLS-B is discussed in further detail in the Specification of the Model section.

The ECLS-K focuses on children's early school experiences beginning in kindergarten and through middle school. The sample size for the combined file (kindergarten to eighth grade) contains 21,409 observations. The ECLS-K data provide descriptive information on children's status at entry to school, transition into school, and progression through eighth grade. The longitudinal nature of the ECLS-K data enables researchers to study how a wide range of family, school, community, and individual factors are associated with school performance. The ECLS-K is a longitudinal study that followed the same children from kindergarten through the eighth grade, so older ages of the ECLS-K sample overlap with the youngest ages of the NLSY97 sample. Information was collected in the fall and the spring of kindergarten (1998–99), fall and spring of first grade (1999–2000), spring of third grade (2002), the spring of fifth grade (2004), and spring of eighth grade (2007). Children, their families, teachers, and schools provided information on children's cognitive, social, emotional, and physical development. Information on their home environment, home educational activities, school environment, classroom environment, classroom curriculum, and teacher qualifications also were collected.[3]

The NLSY97 is a nationally representative survey that gathers information on youth between the ages of 12 and 18 in 1997 and follows them over time. The NLSY97 asks questions annually from 1997 through 2011, and then biannually from 2011 on. Our model uses data from 1997–2011, 2013, and 2015. The starting sample size is 8,984. The survey asks about the youths' family, friends, behavior, education, and economic circumstances. We selected variables from these two data sets that provide important measures of well-being. Details of the variables are outlined in the following sections.

Table 1 shows which dataset provided the main model variables for each life stage. Main model variables are the key outcome variables we track through each stage of a child's development. We supplement the main model variables with context variables that provide information on a child's

family, school, and neighborhood. If a variable was missing for an individual at the point in time of interest, we looked to see if that variable was available in an adjacent previous life stage (ECLS-K) or year (NLSY97) and used that value instead. We show the life stage or age used for this "nearest neighbor" imputation in the table. In the NLSY dataset, if the variable for one age was missing, we took the value from the prior year. If that was also not available, we took the value from the following year. For instance, if a respondent's age 19 response was missing, we looked at their age 18 response. If that was also missing, we went to age 20, and so on.

TABLE 1

**Sources of Data by Life Stage for the Social Genome Model**

| Life stage | Dataset | Main survey /ages used | Imputed survey/ages used (If needed) |
|---|---|---|---|
| Circumstances at Birth (CAB) | | | |
| | NLSY | CAB circumstances reported at age 15 | CAB circumstances reported at age 12–18 |
| | ECLS-B | 9 months | NA |
| Early Childhood (EC) | ECLS-B | 2 years old | NA |
| Preschool (Pre) | | | |
| | ECLS-B | 4 years old | NA |
| | ECLS-K | Kindergarten | NA |
| Early Elementary (Elem) | ECLS-K | 3rd grade | 1st grade |
| Middle Childhood (MC) | ECLS-K | 5th grade | 3rd grade |
| Early Adolescence (EAdol) | NLSY | Age 15 | Age 12–18 |
| Adolescence (Adol) | NLSY | Age 19 | Age 17–21 |
| Transition to Adulthood (TTA) | NLSY | Age 24 | Age 22–26 |
| Adulthood (Adt) | NLSY | Age 30 | Age 28–32 |

**Source:** Social Genome Model.
**Note:** NA = not applicable.

# Matching Protocols

Matching observations from the ECLS-K to the NLSY was a critical step in creating the model. Individuals in the NSLY97 and ECLS-K were both surveyed in early adolescence (around age 15). This period of overlap is what allows us to match the two datasets. The rest of this section describes how we conducted the match.

# Phase 0: Exclusions, Weights, and Imputations

We excluded observations[4] that are missing information on more than 75 percent of key variables that were potential outcome variables in the main model or potential variables to be used to create the matched panel data set. Specifically, we dropped 491 ECLS-K observations because they lacked data for more than 75 percent of key outcome variables. In addition, we excluded observations that are missing information on either race/ethnicity, gender, or both. This excluded 28 additional observations, all from the ECLS-K. We also excluded observations that have an ECLS-K weight of 0 in Kindergarten. This excluded 1,230 additional observations, all from the ECLS-K. No observations were dropped from the NLSY97.

For the matching process, we temporarily imputed missing values using a probit regression, so that observations with any remaining missing data were matched based on our best guess of their matching variable value, rather than on a "missing" category. The order of these imputations is described in table A.1.

# Phase 1: Create Dataset Groups and Category Buckets

TABLE 2

**Identified Sample Groups Based on Race/Ethnicity and Sex**

| Group | ECLS-K | NLSY97 |
|---|---|---|
| Male non-Hispanic, non-Black | 6,860 | 2,453 |
| Female non-Hispanic, non-Black | 6,461 | 2,295 |
| Male Hispanic | 1,752 | 977 |
| Female Hispanic | 1,697 | 924 |
| Male non-Hispanic Black | 1,448 | 1,169 |
| Female non-Hispanic Black | 1,442 | 1,166 |
| **Total** | **19,660** | **8,984** |

**Source**: Social Genome Model.
**Note:** ECLS-K = Early Childhood Longitudinal Study, Kindergarten Class of 1998–99; NLSY97 = National Longitudinal Survey of Youth 1997.

Within each gender/race-ethnicity group, we sequentially divided each group by a set of categories to create more specific buckets of observations (e.g., a single bucket could be: male, Hispanic, above average math score, absent less than 10 days, mother's education is some college or associate's degree, above 200 percent of the federal poverty level [FPL]).

We stopped dividing a group when dividing it further would have created a bucket of 10 or fewer observations from a given dataset (typically from the NLSY97, because it is smaller). We used the datasets that included the imputed values for the category variables (developed from the probit methodology above) and passed each race/ethnicity and sex group through each category division. When a group did not pass through a division due to the 10-observation criterion, we tried variants with combined categories (indicated in **bold** in the list below) in the same division before moving to the next category. Given the existing sample sizes of both datasets, it is unsurprising that we end up with fewer buckets for groups of Hispanic and non-Hispanic Black people.

The observations in each sex and race/ethnicity group were divided into buckets using categories in the following order:

1. Math score
    - Categories: Above, below, near average
2. Days absent
    - Categories: Absent 10 days or more, absent less than 10 days
3. Mother's education
    - Categories for first variant: Less than high school, high school degree/GED, some college or associate's degree, bachelor's degree or higher
    - Categories for second variant: **Less than high school or high school degree/GED**, some college or associate's degree, bachelor's degree or higher
    - Categories for third variant: Less than high school, high school degree/GED, **some college or associate's degree or bachelor's degree or higher.**
4. Twice poverty level
    - Categories: Above 200 percent, below 200 percent
5. Mother's age at first birth
    - Categories for first variant: 17 and under, 18–24, 24+
    - Categories for second variant: **24 and under**, 24+
6. Rural or urban
    - Categories: rural, urban

TABLE 3

**Statistical Match Buckets by Sex and Race/Ethnicity**

| | Number of buckets |
|---|---|
| Male non-Hispanic non-Black | 83 |
| Female non-Hispanic non-Black | 86 |
| Male Hispanic | 36 |
| Female Hispanic | 36 |
| Male non-Hispanic Black | 37 |
| Female non-Hispanic Black | 38 |
| **Total** | **316** |

**Source**: Social Genome Model.

# Phase 2: Divide Buckets into Quintiles

In each bucket, we generated five quintiles, within which we matched individual observations. For each race/ethnicity–gender group, we estimated a logit model by category of mother's education, with the dependent variable equal to the likelihood of being in the ELCS-K dataset. Thus overall, we ran 30 separate logit regressions to estimate the predicted probability that a case came from the ELCS-K (six groups multiplied by five education categories: less than high school, high school degree/GED, some college or associate's degree, bachelor's degree or higher, other).

The independent variables in these regressions were: imputed family poverty level, mother's age at respondent's birth, mother's age at first birth, urban/rural, region, and math score. We weighted outcomes based on rescaled weights.[5]

Once we obtained the predicted probability of being in the ELCS-K, we sorted the ELCS-K observations in each bucket into quintiles based on this predicted probability. These quintiles were also weighted based on the rescaled weight.

After the ECLS-K observations were sorted into quintiles (within each gender/race-ethnicity group and bucket), we identified the minimum and maximum value of the propensity score for each quintile. We then assigned NLSY97 quintiles within each group and bucket based on the "border" values of the ELCS-K values for each quintile.

In some instances, there were no NLSY97 predicted probabilities within a given ELCS-K quintile interval. In those cases, we merged the quintile into the next lower quintile when possible. There were 1,409 separate "matching" cells (composed of Group-Bucket-Quintile blocks) on which we conduct individual matching via lottery.

# Phase 3: Run a Lottery Within Matching Cells

Within each of the 1,409 cells, we expanded the dataset by the rescaled integer weight of the individual observation. This process gave observations with more weight more "tickets" in the matching lottery. Because we weighted by ELCS-K observations when building the quintiles, the quintiles are roughly the same size within each Group-Bucket.

Within each cell, we ran a lottery, randomly selecting one ECLS-K and one NLSY97 observation to be paired together. We conducted this lottery with replacement, such that the same observation could be paired multiple times to observations from the other dataset.

The number of lottery draws for each Group-Bucket-Quintile was determined by the total ECLS-K rescaled weight in the current Group-Bucket-Quintile, as a share of the overall ECLS-K sample, multiplied by 100,000. We end up with lottery draws that look like this at the Group level:

TABLE 4

**Lottery Draws by Group**

| | Number of lottery draws |
|---|---|
| Male non-Hispanic non-Black | 33,601 |
| Female non-Hispanic non-Black | 30,978 |
| Male Hispanic | 9,786 |
| Female Hispanic | 9,253 |
| Male non-Hispanic Black | 8,321 |
| Female non-Hispanic Black | 8,071 |

**Source**: Social Genome Model.

# Final Model Data

With this matched dataset, we employed a multivariate imputation by chained equations procedure to fill in missing data in the matched file. On average, 16 percent of the values for each variable in the final model were imputed. Following this imputation, we ended up with a model with just over 400,000 observations (each matched observation imputed four times). For each observation, we have data on a set of "main model" variables at each life stage. The means of each variable, broken out by race/ethnicity and gender, can be seen in table A.2. The model also includes a number of contextual variables, which are not included in the summary statistics below but are discussed in the section on specifying the model.

As prescribed by our theoretical framework, the main model variables fit roughly into five domains: cognitive and academic development, emotional/psychological development and mental health, physical health and safety, and social behaviors. In the early life stages of the model, we have variables for each domain. Within each domain, the measures vary by life stage, reflecting the ages of the respondents. As we move later in the life course, we pare down the numbers of variables and domains. This section describes the variables measured as part of the main model at each life stage. Below is the list of main model variables, broken out by life stage:

- Circumstances at birth
    - Birth weight
    - Parents married at birth
    - Mother with high school degree
    - Mother with some college or associate's degree
    - Mother with bachelor's degree
    - Mother's age at first birth
- Early childhood[6] (2 years old)
    - Secure toddler attachment
    - General mental ability
    - Overall health status of child
- Preschool (Kindergarten)
    - Math score
    - Reading school
    - Internalizing behavior

- Externalizing behavior
- Parent-child relationship
- Interpersonal skills
- Self-control
- Health

- Early elementary school (3rd grade)
  - Math score
  - Reading score
  - Internalizing behavior
  - Externalizing behavior
  - Parent-child relationship
  - Self-control
  - Health

- Middle childhood (5th grade)
  - Math score
  - Reading score
  - Internalizing behavior
  - Externalizing behavior
  - Peer relationships
  - Self-control
  - Health

- Early adolescence (15 years old)
  - ASVAB score
  - PIAT math score
  - Delinquency index
  - Positive peer behavior
  - Negative peer behavior
  - Mental health
  - Arrested by early adolescence
  - Health
  - Absent from school
  - Suspended for 6+ days

- Adolescence (19 years old)
    - Received high school degree
    - GPA
    - Delinquency index
    - Asks mother and/or father advice
    - Had a child by adolescence
    - Mental health
    - Health
    - Suspended for 6+ days
    - Convicted of or plead guilty to crime
- Transition to adulthood (24 years old)
    - Income-to-poverty ratio
    - Drank before work or school
    - Receiving income from job
    - Not in poverty with a child
    - Mental health
    - Health
    - Convicted of or plead guilty to crime
    - Received high school degree
    - Received associate's degree
    - Received bachelor's degree
    - Received 30 credits or more of higher education, but no degree
    - Completed training or certificate program
    - Inflation-adjusted income
- Adulthood (30 years old)
    - Income-to-poverty ratio
    - Drank before work or school
    - Receiving income from job
    - Not in poverty with a child
    - Mental health
    - Health
    - Convicted of or plead guilty to crime

- Received associate degree
- Received bachelor's degree
- Received 30 credits or more of higher education, but no degree
- Completed training or certificate program
- Inflation-adjusted income

# Benchmarks and Validation

## Validation of Associations

To check whether the life cycle patterns we created when we merged data from the ECLS-K and NLSY97 resemble patterns observed in actual longitudinal data on youth, we compared some key relationships over time in our merged data with data from the NLSY-79 Child and Young Adult cohort (CNLSY).

Specifically, we looked at relationships between selected early elementary variables and variables in later life stages, then compare those relationships to those we observe in the CNLSY. The CNLSY is an older longitudinal dataset that follows the children born to the female respondents of the NLSY-79 survey. Importantly, we intentionally chose variables from the SGM on both sides of the "data seam" between the ECLS-K and NLSY97 data. Thus, our comparison provides information about the quality of the statistical match. The comparisons are not perfect, and, given that the mothers of the children in the CNSLY had to be living in the US in 1979, there are relatively few observations in the CNLSY for Hispanic, Asian, and other groups who made up a larger share of the US population in the years that the ECLS-K and NLSY97 samples were selected. Nevertheless, these data provide some sense of the magnitude of the effect we should expect. Please note that the ECLS-K-NLSY97 relationships shown below were conducted prior to our multiple imputation step. The table with the comparisons can be found below.

TABLE 5

**Comparison of Correlations in the SGM Matched Panel Data with Data from the CNLSY**

| | CNLSY | | | SGM | | |
|---|---|---|---|---|---|---|
| | **Relationship** | **No covars** | **Covars (Birth year cohort FE)** | **Relationship** | **No covars** | **Covars (Mother's education and age, poverty level, region, urban/rural)** |
| Hispanic female | HS Diploma BbTTA on Eadol Math Score | .63 Sig. | .69 Sig. | HS Diploma by TTA on Elem Math Score | .36 Sig. | .09 Sig. |
| Non-Hispanic Black female | | .37 Sig. | .36 Sig. | | .28 Sig. | .12 Sig. |
| Non-Hispanic non-Black female | | .57 Sig. | .56 Sig. | | .61 Sig. | .32 Sig. |
| Hispanic male | | .17 Not Sig. | .16 Not Sig. | | .31 Sig. | .19 Sig. |
| Non-Hispanic Black male | | .35 Sig. | .36 Sig. | | .31 Sig. | .19 Sig. |
| Non-Hispanic non-Black male | | .55 Sig. | .55 Sig. | | .49 Sig. | .23 Sig. |
| Hispanic female | HS diploma by TTA on Eadol Reading Score | .48 Sig. | .54 Sig. | HS Diploma by TTA on Elem Reading Score | .31 Sig. | -.02 Not Sig. |
| Non-Hispanic Black female | | .37 Sig. | .37 Sig. | | .18 Sig. | -.02 Not Sig. |
| Non-Hispanic non-Black female | | .41 Sig. | .44 Sig. | | .6 Sig. | .29 Sig. |
| Hispanic male | | .05 Not Sig. | .09 Not Sig. | | .36 Sig. | .23 Sig. |
| Non-Hispanic Black male | | .42 Sig. | .46 Sig. | | .3 Sig. | .18 Sig. |
| Non-Hispanic non-Black male | | .36 Sig. | .44 Sig. | | .46 Sig. | .2 Sig. |
| Hispanic female | Ever Convicted by Adt on Eadol Externalizing Behavior | .15 Not Sig. | .17 Not Sig. | Ever Convicted by Adt on Elem Externalizing Behavior | .16 Sig. | .13 Sig. |
| Non-Hispanic Black female | | .14 Not Sig. | .15 Not Sig. | | -.01 Not Sig. | -.11 Sig. |
| Non-Hispanic non-Black female | | .21 Not Sig. | .22 Not Sig. | | .06 Sig. | -.06 Sig. |
| Hispanic male | | -.04 Not Sig. | -.03 Not Sig. | | .1 Sig. | .11 Sig. |
| Non-Hispanic Black male | | .16 Not Sig. | .19 Not Sig. | | .05 Sig. | -.03 Not Sig. |
| Non-Hispanic non-Black male | | .38 Sig. | .42 Sig. | | .07 Sig. | .02 Not Sig. |

**Source:** Social Genome Model

**Notes:** FE = fixed effects; Sig. = statistically significant; Not sig. = not statistically significant.

The far-left column identifies for which sex and race/ethnicity group the comparison is being done. The first set of columns depicts the relationship between variables in the CNLSY, while the second set of columns depicts the relationship between variables in the SGM. The first column in each set identifies the relationship being compared. For example, the first comparison is a regression of a binary indicator of receiving a high school diploma by the transition to adulthood life stage on math score. In the CNLSY, the math score comes from the early adolescence life stage; in the SGM, math score comes from the elementary life stage. The "no covars" column shows the results of a simple logistic regression with only the previously identified variables (i.e., without covariates). Green text indicates a positive relationship, and red text indicates a negative relationship. Highlighting indicates statistical significance at the 5 percent level. Finally, the "covars" column shows the same regression but with additional independent variables (i.e., with covariates). For the CNLSY, the only additional variables were birth year cohort fixed effects (FE). For the SGM, rather than a birth year cohort fixed effect, we added indicators of mother's education, mother's age at first birth, whether the individual had income above 200 percent of the poverty level, region, and whether the individual lived in an urban or rural area.

Qualitatively speaking, the relationships in our model look similar to those in the CNLSY, especially when looking at the regressions without covariates. Once we add in covariates, the relationships are generally not as strong in our model as they are in the CNLSY. To some degree, this should be expected, given that we included more covariates in the regressions for our model than in the regressions for the CNLSY. Because some of the variables in our matched panel are imputed, there is likely more measurement error in our matched panel than in the CNLSY; measurement error tends to attenuate estimated relationships. We believe this rough comparison shows that our match creates a reasonable longitudinal dataset.

# Summary Statistics

In the table below, we present summary statistics from our final matched dataset. As noted above, the dataset is weighted by sex and race/ethnicity to represent the birth cohort in the year 2000. Our data show that approximately 73 percent of our sample had a high school degree by the time they were 19 years old. According to data from NCES, the public high school graduation rate in 2000 was 72 percent (Kena et al. 2014). In our dataset, the mean age of respondents' mothers at first birth is 23.1. According to CDC data, in 2000, the average age of mothers at first birth was 24.9. Given that our

NLSY97 cohort was generally born in early to middle 1980s, it makes sense that mother's age at first birth would be slightly lower than what it was in 2000 (Matthews and Hamilton 2016).

TABLE 6

**Summary Statistics of Final Dataset**

| | Mean | Standard deviation | Minimum | Maximum |
|---|---|---|---|---|
| Share of observations that are non-Hispanic Black | 0.15 | 0.36 | 0 | 1 |
| Share of observations that are Hispanic | 0.20 | 0.40 | 0 | 1 |
| Share of observations that are non-Hispanic non-Black | 0.65 | 0.48 | 0 | 1 |
| Share of observations that are female | 0.49 | 0.50 | 0 | 1 |
| Birth weight (lbs) | 7.36 | 1.35 | 1 | 13.69 |
| Share of parents married at Birth | 0.66 | 0.47 | 0 | 1 |
| Share of mothers whose highest degree is a high school degree | 0.27 | 0.44 | 0 | 1 |
| Share of mothers whose highest degree is an associate degree | 0.34 | 0.48 | 0 | 1 |
| Share of mothers whose highest degree is a Bachelor's degree | 0.24 | 0.43 | 0 | 1 |
| Mother's age at first birth (tears) | 23.12 | 5.30 | 1 | 63 |
| Share of observations with high school degree by adolescence | 0.73 | 0.44 | 0 | 1 |
| Annual individual earnings in adulthood (positive only) * | $41,931 | $31,275 | $3 | $224,753 |

**Notes**: Monetary values are adjusted for inflation to 2018 dollars using the CPI-U.

*Imputed values for earnings can be negative and those negative imputations are used below for model specification as they preserve the linear relationship between earnings and the variables included in the mode.

The ECLS-K and NLSY measures included in the SGM differ from other national datasets, and this makes it impossible to identify an exact benchmark comparison. However, we have examined numerous data sources and we consistently find similar patterns in adult outcomes between the SGM and nationally representative data collected by the Census Bureau (Current Population Survey (CPS), Annual Social and Economic Supplements) and the Centers for Disease Control (National Health Interview Survey (NHIS); Behavioral Risk Factor Surveillance System (BRFSS) within and across the six race/ethnicity and sex groups. The comparable patterns across datasets provide a level of validation to the model. For example, we find the following patterns in adult outcomes[7]:

- Earnings: In both the SGM and CPS data, males have higher incomes than females (based on the CPS median income and the mean annual individual earnings in the SGM).
- Education: Within the SGM, a larger proportion of females than males earn an associate degree or a bachelor's degree. These patterns are also found in data from the CPS Annual Social and Economic Supplement, where females reported higher levels of some college (less

than 4-year degree) and bachelor's degree or higher than their male counterparts in the three race/ethnicity groups.

    - The SGM differs from the national data when broken down by race/ethnicity. In the SGM, Hispanic females have higher levels of education than Black females; 2018 CPS data indicates that education levels are higher among Black females than females with Hispanic or Latino origin.
    - In the CPS data, a greater proportion of white females are receiving degrees than Black and Hispanic/Latino females. In the SGM, a larger proportion of the non-Black, non-Hispanic female group have degrees; however, as noted, the non-Black, non-Hispanic groups in the SGM include adults who are white as well as those who identify as another race.

- Employment: The SGM has a proxy measure for employment with a binary variable for whether an individual is receiving pay from a job, whereas the CPS has three categories of employment: (1) employed, (2) unemployed, and (3) caring for children, armed forces, or not in labor force. Though these measures are not entirely comparable to the SGM measure, the patterns we see in the SGM and CPS data are similar:

    - A higher percentage of males than females are receiving pay from a job (SGM)/are employed (CPS) than females.
    - Among males, Hispanic males have the highest percentage of adults who are receiving pay from a job (SGM)/employed (CPS).

- Mental Health: Data from the NHIS and the SGM both indicate that females of all race/ethnicity groups have poorer mental health compared to their male counterparts.
- Health: Data from the NHIS and BRFSS indicate that the majority of U.S. adults are in good, very good, or excellent health—a pattern we see in the SGM as well.

    - In both the NHIS, BRFSS and SGM data, males report good health at slightly higher rates than females,
    - In the NHIS and BRFSS Hispanic adults have the poorest health outcomes, with Black and white adults reporting better health outcomes. The same pattern is true in the SGM data, however, the non-Black, non-Hispanic racial and ethnic group includes both white and Asian adults, so is not directly comparable to the "white only" data in the national data sources.

Exact sources for comparison:

Mental Health: "Table A-7a, Age-adjusted percentages (with standard errors) of feelings of sadness, hopelessness, worthlessness, and that everything is an effort, among adults aged 18 and over, by selected characteristics: United States, 2018," National Health Interview Survey 2018, accessed January 26, 2021, https://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2018_SHS_Table_A-7.pdf; and "Table A-8a, Age-adjusted percentages (with standard errors) of feelings of nervousness, feelings of restlessness, and serious psychological distress among adults aged 18 and over, by selected characteristics: United States, 2018," National Health Interview Survey 2018, accessed January 26, 2021, https://ftp.cdc.gov/pub/Health_Statistics/NCHS/NHIS/SHS/2018_SHS_Table_A-8.pdf.

- Health: KFF analysis of the Centers for Disease Control and Prevention (CDC)'s 2019 Behavioral Risk Factor Surveillance System (BRFSS). See "Males Who Report Fair or Poor Health Status, by Race/Ethnicity, 2019," Kaiser Family Foundation, accessed January 26, 2021, https://www.kff.org/racial-equity-and-health-policy/state-indicator/male-self-reported-fair-or-poor-health-status-by-raceethnicity/; and "Females Who Report Fair or Poor Health Status, by Race/Ethnicity, 2019," Kaiser Family Foundation, accessed January 26, 2021, https://www.kff.org/racial-equity-and-health-policy/state-indicator/female-self-reported-fair-or-poor-health-status-by-raceethnicity/
- Education: US Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2018. Estimates of Adult Civilian Persons. Tables were generated using the US Census Bureau's CPS Table Creator at https://www.census.gov/cps/data/cpstablecreator.html.
- Employment: US Census Bureau, Current Population Survey, Annual Social and Economic Supplement, 2018. Estimates of Adult Civilian Persons. Tables were generated using the US Census Bureau's CPS Table Creator at https://www.census.gov/cps/data/cpstablecreator.html.
- Earnings: US Census Bureau, Current Population Survey, 2017 and 2018 Annual Social and Economic Supplements. See Fontenot, Semega, and Collar (2018, table 1).

# Specification of the Model

We specified the Social Genome Model using an iterative algorithm, testing the importance of each main model variable from earlier life stages, as well as contextual variables in the current life stage. All main model variables, and some specific late life stage variables, have the chance to influence the output in the modeled life stage. Context variables can also affect the life stage, but these effects do not carry into the next stage, except through the main model variables. We ran the model separately for each gender and race/ethnicity group. Below is the list of context variables in the model that come from the ECLS-K and NLSY, broken out by life stage.

## Context Variables

- Preschool (Kindergarten)
    - Attended prekindergarten the year before kindergarten
    - Attended Head Start program the year before kindergarten
    - Received non-preschool/prekindergarten/Head Start center-based care the year before kindergarten
    - Received non-relative care the year before kindergarten
- Early elementary school (3rd grade)
    - Child obesity
    - Parent school involvement
    - Teacher turnover
    - Father in household
    - Out-of-school activities
    - Positive stimulation
    - Routines
    - SNAP/food stamps
    - Household income-to-poverty ratio
    - No health insurance
    - Parental support
    - Neighborhood safety
    - Neighborhood issues

- Household income-to-poverty ratio

Middle childhood (5th grade)

- Hearing and seeing problems
- Child obesity
- Teacher turnover
- Father in household
- Parent school involvement
- Out-of-school activities
- Positive stimulation
- Parental support
- Routines
- SNAP/food stamps)
- Household income-to-poverty ratio
- No health insurance
- Family home ownership
- Family food insecurity status
- Household income-to-poverty ratio
- Neighborhood safety
- Neighborhood issues
- Negative discipline

- Early adolescence (15 years old)
  - Authoritative parent
  - Father in household
  - Gangs in school or neighborhood
  - Family net worth

- Adolescence (19 years old)
  - Authoritative parent
  - Father in household
  - Gangs in school or neighborhood
  - Victim of a violent crime
  - Lives in rural area

- Transition to adulthood (24 years old)
    - Lives in a rural area
    - Limited in amount or kind of work
- Adulthood (30 years old)
    - Lives in a rural area
    - Limited in amount or kind of work

# Estimation Procedure

We ran a set of ordinary-least squares regressions, sequentially from earliest life stage to latest life stage. For example, we first ran the preschool life stage with circumstances at birth variables and preschool context variables, then we ran early elementary with circumstances at birth variables, preschool main model variables, and early elementary context variables. Within each life stage, we estimated an equation for each main model variable. This process is referred to as "parameterization."

All of the main model variables were constructed in a way such that the expected coefficient would be positive. For example, "Convicted or Pled Guilty to Crime" is reversed in the model specification (0, 1 becomes -1, 0), such that the expected coefficient between the convicted guilty variable and an outcome like bachelor's degree attainment is positive (i.e., those who are not convicted of a crime are more likely to get a degree). This allowed us to easily identify the variables with negative coefficients as those that were eligible to be "pruned" from the model.

Main model variables were tested for two criteria: expected sign and measure of goodness of fit (adjusted R-squared). Main model variables were always retained if the sign of the coefficient was consistent with theory (positive). In some cases, a main model variable may have a negative, or unexpected, coefficient sign, but still be integral in explaining the outcome of a model. To test for this, when we excluded a main model variable, we assessed if the adjusted R-squared value decreased by more than 0.015 (1.5 percentage points). If the adjusted R-squared, or goodness-of-fit value, decreased by more than this amount, we took this to mean that the variable was a key explainer of the specified outcome, even though the coefficient was in an unexpected direction. Context variables were only tested with an adjusted R-square threshold. If removing a context variable decreased the adjusted R-squared value by more than 0.005 (0.5 percentage points), the context variable was retained in the model.

We continued to test the main model variables using a similar iterative process. We tested the most "distant" variables first, both in terms of time and domain relation. For example, if the variable was standardized reading score (domain 1) in the early elementary (Elem) life stage, we first tested context variables using the process above, then tested circumstances at birth (CAB) variables, then preschool (Pre) variables. Because the Pre variables aligned with the five domains, we tested the farthest variable first (physical health in Pre, domain 5), then moved towards the closest domain. Thus, we next tested interpersonal skills in Pre (domain 4), parent-child relationship (domain 3), internalizing and externalizing behaviors (domain 2) and math and reading scores (domain 1).

When all variables were either removed, had the "expected" coefficient sign, or were retained in the model due to meeting the adjusted R-squared threshold, the resulting coefficients became the parameterization metrics for the given outcome and gender and race/ethnicity group, along with the constant, and the individual-level residuals from the final regression predicting each outcome were saved within the dataset.

This process ensures that each life stage could, in theory, affect outcomes in each subsequent life stage, along with contextual variables. There are two exceptions to this rule. First, in this version of the model, we can only provide coefficients describing the relationship between early childhood variables and preschool main model variables for the early childhood (EC) life stage. Thus, an intervention in the EC stage only directly affects outcomes in the preschool life stage, which then flow through the remainder of the model. Second, we remove mother's education and mother's age at first birth from the regressions that predict the early adolescent variables. We remove the variables from these regressions at this stage because the early adolescent life stage is where we knit together data from the ECLS-K and NLSY97, creating a seam in the date. Mother's education and mother's age at first birth were used in the matching process, so they tend to attenuate the effects of the other variables in the regressions at that life stage.

# Early Childhood

For the EC life stage, we use the ECLS-B. The ECLS-B is a nationally representative survey of children born in the year 2001. The ECLS-B is a "multisource, multimethod study that focuses on the early home and educational experiences of children during their first 6 years."[8] The study is sponsored by the National Center for Education Statistics (NCES), located within US Department of Education and the Institute of Education Sciences, in collaboration with several federal education and health policy agencies. As a longitudinal study, the same children were followed from birth through kindergarten

entry. Data were obtained from birth certificates, nine-month surveys, and assessments from when the children were approximately 9 months old (2001–2002), 2 years old (2003–2004), 4 years old/preschool age (2005–2006), and kindergarten age (fall of 2006, fall of 2007). Children, their parents, their child care and early education providers, and their teachers provided information on children's cognitive, social, emotional, and physical development across multiple settings (e.g., home, child care, school)[9]

The ECLS-B is a restricted-use dataset that can only be used by licensed users in a secure room that does not have access to the internet or through restricted remote access. This severely limits its use in the context of the SGM. Child Trends and the Urban Institute explored a variety of alternatives to address this issue and enable the inclusion of ECLS-B data into the model, but none were able to ensure compliance with NCES security procedures for the data, while allowing the team to use the data in an unrestricted manner. As a consequence, in this version of the model, the ECLS-B data are not matched with the ECLS-K or the NLSY. Rather, we used the restricted-use ECLS-B to run regressions, and the coefficients from those regressions to estimate the impact of changes in early childhood on select pre-kindergarten outcomes. Specifically, we regressed preschool outcomes that are common to both the ECLS-B and the ECLS-K (math scores, reading scores, and overall health of the child) on the CAB variables, contextual variables, and EC variables (secure toddler attachment, general mental ability, and overall health status of child).

The coefficients from these regressions are included in the SGM, but the underlying data are not. This approach enables the SGM to estimate the effects of early childhood interventions by changing math, reading, and health in the ECLS-K in magnitudes equivalent to the coefficients obtained in the regressions using the ECLS-B. However, this approach means that EC variables can only affect later life outcomes through their effect on one of the three Preschool stage variables, excluding any other potential paths (e.g., changes in an early childhood variable that affect internalizing behaviors in Preschool). Also, because the actual individual level data are not included in the model, we cannot use any EC characteristics to select subpopulations for simulations (e.g., those in families with income below 200 percent of the FPL at age 2).

A second version of the model uses the underlying ECLS-B data, which allows the model to have additional early childhood variables, and paths between early child variables and variables at later life stages. This version can only be accessed through a restricted data license with NCES. More information about this version can be found on the Child Trends website.

# Process for Simulating Outcomes

When we want to simulate the effect of a program, a policy intervention, or "what-if" scenario in the Social Genome Model, we can change a given main model variable by the size and direction as documented by an outside study of an intervention or by an aspirational amount. For example, we might want to look at the effect of increasing reading scores in middle childhood for children in families with incomes below 200 percent of the FPL. In this simple example, we increase each eligible child's reading score by a given amount. We could also assign an effect using a normalized distribution so the average effect for all treated children reaches the given amount while any individual child may benefit more than or less than average, or cap the effect for students who already have high reading scores. For binary variables in the model, such as bachelor's degree attainment, we implement interventions by increasing the share of observations meeting the criteria. For example, we randomly provide bachelor's degrees to individuals who do not have them until we reach the intervention level (for example, increasing the percent of people in a given population who have bachelor's degrees by 3 percentage points).

When we simulate an intervention, like increasing reading scores in middle childhood, we assign the higher reading scores to the treated individuals, and then use the coefficients from parametrization to predict a new value for outcomes that have reading scores as dependent variable, and then use those new outcomes to predict subsequent outcomes. For example, if a higher reading score in middle childhood is associated, in our model, with better mental health in early adolescence, then treated individuals would see a slight boost in their mental health scores, above their previous levels. In the next life stage, adolescence, outcomes could potentially be affected not only by the direct intervention (increased reading score in middle childhood), but also by secondary effects (improved mental health). In this way, primary and secondary effects could have an influence in the model, all the way to adulthood.

When interventions are applied in the early childhood life stage in this version of the model, we apply the intervention manually, as we do not have underlying data for our ECLS-K/NLSY97 dataset. For example, all observations have a mental ability value of 0 (mean standard deviation value) in the model. An intervention that increases mental ability by 0.2 standard deviation is applied by adding the value to this mean of 0. We then apply the coefficients on mental ability to predict Preschool reading, math, and mental health (the only three preschool stage main model variables that can be directly affected by changes in EC factors, and the subsequent stages of the model continue as they would for any other intervention.

# Example Application

To demonstrate a simulation, we present an example of an application. In this example, we increased math and reading scores in preschool by one standard deviation for all observations in the model. This increase in test scores is "aspirational," meaning that we do not have a specific intervention in mind that would cause a one standard deviation in reading and math test scores for all children in the United States. Other simulations could be run to adjust reading and math scores by an amount that has been found by a randomized control trial evaluation of a real-life intervention.

We present the results of this simulation below. These results focus on income and degree attainment for simplicity's sake, but users of the SGM can see how their interventions affected all of the variables in the model at each life stage subsequent to the intervention (i.e., everything after Preschool).

TABLE 7

**Results of Example Intervention**

| | Prior to intervention | After intervention | Change |
|---|---|---|---|
| Received high school degree by adolescence | 73.2% | 74.3% | 1.1% |
| Received associate's degree by adulthood | 10.8% | 11.2% | 0.4% |
| Received bachelor's degree by adulthood | 26.8% | 29.7% | 2.9% |
| Inflation-adjusted income in adulthood | $33,605 | $35,961 | $2,356 |
| Lifetime earnings | $653,903 | $692,966 | $39,062 |

**Source:** Social Genome Model.

We see improvements across a variety of measures from this increase in math and reading scores. It is notable that an intervention so early in life, in this case when the individual is just four years old, has a visible effect many years down the road.

# Appendix

TABLE A.1

**Imputations from Matching Process**

*ECLS-K*

| Variable imputed | Using variables | Observations imputed |
|---|---|---|
| Mother's education | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, math score, biological dad at home | 396 |
| Poverty level (age 15) | Urban/rural, gender, race/ethnicity, region, mother's age, math score, mother's education | 175 |
| Mother's age at first birth | Poverty level, urban/rural, gender, race/ethnicity, region, mother's education, math score, biological dad at home | 3,210 |
| Math score (age 15) | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, mother's education, class size | 9,225 |
| Absences (age 15) | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, mother's education, parents involved, suspensions, health, repeat grade | 11,008 |

TABLE A.2

**Imputations from Matching Process**

*NLSY97*

| Variable imputed | Using variables | Observations imputed |
|---|---|---|
| Absences (age 15) | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, math score | 297 |
| Mother's education | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, math score, biological dad at home | 503 |
| Mother's age at first birth | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, math score, biological dad at home | 717 |
| Poverty level (age 15) | Urban/rural, gender, race/ethnicity, region, mother's age, mother's education, math score, receive AFDC/TANF, gangs in school/neighborhood | 922 |
| Math score (p1) (age 15) | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, mother's education, ASVAB score | 1,735 |
| Math score (p2) (age 15) | Poverty level, urban/rural, gender, race/ethnicity, region, mother's age, mother's education | 1,105 |

TABLE A.3

**Means of Main Model Variables by Race/Ethnicity and Sex**

| | Black female | Black male | Hispanic female | Hispanic male | Non-Black, Non-Hispanic female | Non-Black, Non-Hispanic male |
|---|---|---|---|---|---|---|
| **Circumstances at birth** | | | | | | |
| Birth weight (Cont.) | 6.74 | 7.06 | 7.18 | 7.44 | 7.33 | 7.61 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Were parents married at birth? (Prop.) | 0.27 | 0.27 | 0.59 | 0.59 | 0.78 | 0.78 |
| Does mother have at least a high school degree or GED? (Prop.) | 0.33 | 0.32 | 0.28 | 0.26 | 0.26 | 0.25 |
| Does mother have at least some college? (Prop.) | 0.39 | 0.38 | 0.30 | 0.30 | 0.36 | 0.35 |
| Does mother have at least a bachelor's degree? (Prop.) | 0.12 | 0.12 | 0.11 | 0.11 | 0.30 | 0.30 |
| Mother's age at first birth (Cont.) | 20.47 | 20.54 | 21.64 | 21.60 | 24.33 | 24.34 |
| **Preschool** | | | | | | |
| Math score (SD) | -0.39 | -0.45 | -0.41 | -0.48 | 0.22 | 0.24 |
| Reading score (SD) | -0.23 | -0.40 | -0.28 | -0.56 | 0.24 | 0.05 |
| Internalizing behavior (SD) | -0.04 | -0.11 | -0.02 | -0.05 | 0.07 | -0.01 |
| Externalizing behavior (SD) | -0.07 | -0.54 | 0.24 | -0.13 | 0.28 | -0.14 |
| Parent-child relationship (SD) | 0.11 | 0.05 | 0.08 | -0.02 | 0.00 | -0.05 |
| Interpersonal skills (SD) | -0.10 | -0.43 | 0.08 | -0.24 | 0.28 | -0.08 |
| Self-control (SD) | -0.15 | -0.49 | 0.09 | -0.17 | 0.28 | -0.08 |
| Health (SD) | -0.17 | -0.26 | -0.16 | -0.30 | 0.13 | 0.08 |
| **Elementary (3rd grade)** | | | | | | |
| Math score (SD) | -0.59 | -0.52 | -0.35 | -0.29 | 0.10 | 0.27 |
| Reading score (SD) | -0.36 | -0.56 | -0.26 | -0.50 | 0.27 | 0.10 |
| Internalizing behavior (SD) | -0.02 | -0.17 | 0.01 | -0.03 | 0.02 | -0.04 |
| Externalizing behavior (SD) | -0.17 | -0.67 | 0.25 | -0.15 | 0.26 | -0.14 |
| Parent-child relationship (SD) | 0.02 | 0.07 | 0.04 | 0.04 | 0.00 | -0.05 |
| Self-control (SD) | -0.23 | -0.58 | 0.17 | -0.18 | 0.24 | -0.11 |
| Health (SD) | -0.22 | -0.25 | -0.26 | -0.28 | 0.13 | 0.08 |
| **Middle childhood (5th grade)** | | | | | | |
| Math score (SD) | -0.66 | -0.61 | -0.40 | -0.25 | 0.06 | 0.25 |
| Reading score (SD) | -0.43 | -0.65 | -0.30 | -0.50 | 0.24 | 0.08 |
| Internalizing behavior (SD) | -0.01 | -0.07 | 0.00 | -0.09 | 0.05 | -0.11 |
| Externalizing behavior (SD) | -0.24 | -0.71 | 0.29 | -0.24 | 0.28 | -0.20 |
| Peer relationships (SD) | 0.15 | 0.15 | -0.03 | -0.12 | 0.08 | -0.12 |
| Self-control (SD) | -0.32 | -0.63 | 0.24 | -0.23 | 0.28 | -0.19 |
| Health (SD) | -0.22 | -0.29 | -0.27 | -0.29 | 0.14 | 0.08 |
| **Early adolescence (age 15)** | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| PIAT math score (SD) | -0.58 | -0.53 | -0.21 | -0.16 | 0.13 | 0.26 |
| ASVAB score (SD) | -0.64 | -0.84 | -0.39 | -0.44 | 0.24 | 0.15 |
| Delinquency index (SD) | 0.23 | -0.13 | 0.14 | -0.25 | 0.16 | -0.14 |
| Positive peer behavior (SD) | -0.01 | -0.04 | -0.19 | -0.22 | 0.12 | 0.02 |
| Negative peer behavior (SD) | -0.24 | -0.13 | -0.17 | -0.06 | -0.03 | 0.20 |
| Mental health (SD) | -0.30 | 0.18 | -0.32 | 0.27 | -0.21 | 0.20 |
| Arrested by early adolescence (Prop.) | 0.08 | 0.22 | 0.07 | 0.18 | 0.08 | 0.12 |
| Health (SD) | -0.17 | 0.06 | -0.17 | 0.03 | -0.10 | 0.15 |
| Absent from school (number of days, Cont.) | 5.71 | 5.41 | 6.66 | 6.82 | 6.59 | 5.87 |
| Suspended from school for 6+ days (Prop.) | 0.05 | 0.13 | 0.02 | 0.11 | 0.03 | 0.07 |
| **Adolescence (age 19)** | | | | | | |
| Received high school diploma (Prop.) | 0.73 | 0.56 | 0.71 | 0.63 | 0.80 | 0.74 |
| GPA (SD) | -0.20 | -0.66 | 0.00 | -0.47 | 0.13 | -0.18 |
| Delinquency index (SD) | 0.22 | -0.25 | 0.24 | -0.14 | 0.21 | -0.11 |
| Asks mother and/or father for advice (SD) | 0.08 | -0.12 | 0.12 | -0.07 | 0.20 | -0.20 |
| Had a child by adolescence (Prop.) | 0.35 | 0.17 | 0.26 | 0.11 | 0.18 | 0.07 |
| Mental health (SD) | -0.14 | 0.08 | -0.29 | 0.25 | -0.18 | 0.19 |
| Health (SD) | -0.16 | -0.02 | -0.16 | 0.05 | -0.07 | 0.13 |
| Suspended from school for 6+ days (Prop.) | 0.01 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 |
| Convicted of or plead guilty to crime (Prop.) | 0.07 | 0.21 | 0.05 | 0.18 | 0.08 | 0.17 |
| **Transition to adulthood (age 24)** | | | | | | |
| Income-to-poverty ratio (Cont.) | 244.22 | 243.30 | 347.52 | 311.35 | 406.01 | 405.56 |
| Drank before work or school (Prop.) | 0.09 | 0.12 | 0.07 | 0.10 | 0.07 | 0.05 |
| Receiving income from job (Prop.) | 0.76 | 0.74 | 0.79 | 0.89 | 0.83 | 0.89 |
| Not low income with a child (Prop.) | 0.56 | 0.73 | 0.72 | 0.78 | 0.83 | 0.92 |
| Mental health (SD) | -0.07 | 0.15 | -0.13 | 0.14 | -0.15 | 0.12 |
| Health (SD) | -0.13 | 0.01 | -0.21 | -0.07 | -0.01 | 0.11 |
| Convicted of or plead guilty to crime (Prop.) | 0.12 | 0.35 | 0.11 | 0.28 | 0.12 | 0.25 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Received high school diploma (Prop.) | 0.76 | 0.60 | 0.74 | 0.66 | 0.81 | 0.76 |
| Received associate's degree (Prop.) | 0.05 | 0.03 | 0.10 | 0.04 | 0.08 | 0.07 |
| Received bachelor's degree (Prop.) | 0.11 | 0.06 | 0.13 | 0.08 | 0.31 | 0.23 |
| Received 30 credits or more higher education but no degree (Prop.) | 0.37 | 0.25 | 0.32 | 0.26 | 0.44 | 0.40 |
| Completed training or certificate program (Prop.) | 0.29 | 0.24 | 0.23 | 0.19 | 0.20 | 0.21 |
| Inflation-adjusted earnings (Positive only, Cont.)Cont. | 21,645 | 22,407 | 24,273 | 29,275 | 25,939 | 32,718 |
| Inflation-adjusted earnings (Cont.) | 15,486 | 15,757 | 18,246 | 25,025 | 21,000 | 28,528 |
| **Adulthood (age 30)** | | | | | | |
| Income-to-poverty ratio (Cont.) | 226.74 | 229.19 | 317.07 | 327.95 | 438.86 | 445.28 |
| Drank before work or school (Prop.) | 0.10 | 0.13 | 0.08 | 0.10 | 0.08 | 0.07 |
| Receiving income from job (Prop.) | 0.71 | 0.69 | 0.73 | 0.89 | 0.78 | 0.88 |
| Not low income with a child (Prop.) | 0.52 | 0.64 | 0.68 | 0.77 | 0.78 | 0.87 |
| Mental health (SD) | -0.08 | 0.13 | -0.09 | 0.21 | -0.18 | 0.14 |
| Health (SD) | -0.14 | -0.03 | -0.16 | -0.20 | 0.05 | 0.11 |
| Convicted of or plead guilty to crime (Prop.) | 0.14 | 0.40 | 0.13 | 0.32 | 0.14 | 0.29 |
| Received associate's degree (Prop.) | 0.11 | 0.05 | 0.14 | 0.07 | 0.12 | 0.10 |
| Received bachelor's degree (Prop.) | 0.17 | 0.11 | 0.20 | 0.13 | 0.37 | 0.29 |
| Received 30 credits or more higher education but no degree (Prop.) | 0.39 | 0.30 | 0.34 | 0.29 | 0.42 | 0.40 |
| Completed training or certificate program (Prop.) | 0.39 | 0.34 | 0.34 | 0.28 | 0.30 | 0.32 |
| Inflation-adjusted earnings (Positive only, Cont.) | 27,325 | 29,723 | 32,538 | 42,125 | 39,201 | 51,095 |
| Inflation-adjusted earnings (Cont.) | 18,559 | 19,695 | 23,876 | 36,880 | 30,523 | 44,891 |

**Notes**: Cont. = continuous; prop. = proportion; SD = standard deviation. Monetary values are adjusted for inflation to 2018 using the CPI-U.

TABLE A.4

**Social Genome Model Data Dictionary**

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Birthweight | Birthweight of child (continuous, measured in pounds) | Circumstances at Birth | Circumstances at Birth | ECLS-K |
| Parents married at birth | Biological parents were married when child was born (binary) | Circumstances at Birth | Circumstances at Birth | ECLS-K |
| Mother completed a high school degree or GED | Mother's maximum education received by child's birth is a high school degree or GED (binary) | Circumstances at Birth | Circumstances at Birth | ECLS-K, NLSY |
| Mother completed some college | Mother's maximum education received by child's birth was some college (binary) | Circumstances at Birth | Circumstances at Birth | ECLS-K, NLSY |
| Mother completed college or higher | Mother's maximum education received by child's birth is a bachelor's degree or higher (binary) | Circumstances at Birth | Circumstances at Birth | ECLS-K, NLSY |
| Mother's age at first birth | Mother's age at first birth (continuous) | Circumstances at Birth | Circumstances at Birth | ECLS-K, NLSY |
| Math score | Standardized math proficiency test t-score (continuous) | Preschool | Main Model | ECLS-K |
| Reading score | Standardized reading proficiency test t- score (continuous) | Preschool | Main Model | ECLS-K |
| Internalizing behavior | Standardized internalizing behaviors score SRS (continuous; higher is less internalizing behavior) | Preschool | Main Model | ECLS-K |
| Externalizing behavior | Standardized externalizing behaviors score SRS score (continuous; higher is less externalizing behavior) | Preschool | Main Model | ECLS-K |
| Parent-child relationship | Standardized scale of responses to four items regarding parental relationship with child: have warm, close time together; child likes them; parent shows love; express affection to child. Original scale:1 = worst, 4 = best (continuous) | Preschool | Main Model | ECLS-K |
| Interpersonal skills | Standardized interpersonal skills score SRS (continuous) | Preschool | Main Model | ECLS-K |
| Self-control | Standardized self-control score SRS (continuous) | Preschool | Main Model | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Health | Standardized scale of overall health status of child. Original scale: 1 = excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Preschool | Main Model | ECLS-K |
| Attended preschool/pre-K the year before kindergarten | Child attended preschool/pre-K the year before kindergarten (binary) | Preschool | Contextual | ECLS-K |
| Attended Head Start program the year before kindergarten | Child attended Head Start program the year before kindergarten (proportion) | Preschool | Contextual | ECLS-K |
| Received non-preschool/pre-K/Head Start center-based care the year before kindergarten | Child received non-preschool/pre-K/Head Start center-based care the year before kindergarten (proportion) | Preschool | Contextual | ECLS-K |
| Received non-relative care the year before kindergarten | Child received non-relative care the year before kindergarten (proportion) | Preschool | Contextual | ECLS-K |
| Math score | Standardized math proficiency test t-score (continuous) | Elementary | Main Model | ECLS-K |
| Reading score | Standardized reading proficiency test t- score (continuous) | Elementary | Main Model | ECLS-K |
| Internalizing behavior | Standardized internalizing behaviors score SRS (continuous; higher is less internalizing behavior) | Elementary | Main Model | ECLS-K |
| Externalizing behavior | Standardized externalizing behaviors score SRS score (continuous; higher is less externalizing behavior) | Elementary | Main Model | ECLS-K |
| Parent-child relationship | Standardized scale of responses to four items regarding parental relationship with child: have warm, close time together; child likes them; parent shows love; express affection to child. Original scale:1 = worst, 4 = best (continuous) | Elementary | Main Model | ECLS-K |
| Self-control | Standardized self-control score SRS (continuous) | Elementary | Main Model | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Health | Standardized scale of overall health status of child. Original scale:1 = excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Elementary | Main Model | ECLS-K |
| Child obesity | Child measured in 95th percentile based on CDC growth chart and is obese (binary) | Elementary | Contextual | ECLS-K |
| Parent school involvement | Index of responses to items related to parent-school involvement: open houses, PTA meetings, parent-teacher conferences, school events, volunteers, fundraising (continuous, 0–1) | Elementary | Contextual | ECLS-K |
| Teacher turnover | Teacher turnover is a problem at this school, as reported by school administrator (binary) | Elementary | Contextual | ECLS-K |
| Biological father in the home | Presence of biological father in the home (binary) | Elementary | Contextual | ECLS-K |
| Out-of-school activities | Child has ever participated in one or more out-of-school activities: dance, athletics, clubs/recreation programs, music lessons, art classes, performing arts (binary) | Elementary | Contextual | ECLS-K |
| Positive stimulation | Index of responses to items regarding positive stimulation activities: (1) number of children's books in home; (2) outings to library, museum, concert, zoo, and/or sporting event in the past month; (3) positive stimulation activities done with child in a typical week (stories, singing, arts/crafts, involved in chores, games/puzzles, talk about nature/science, building, sport/exercise, practice reading/writing/numbers, read to child (continuous, 0–3) | Elementary | Contextual | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Routines | Index of responses to three items regarding household rules and routines: (1) rules around television (which programs, how early/late, hours during weekdays); (2) regular bedtime; (3) number of meals per week family eats breakfast and dinner together (continuous , 0–3) | Elementary | Contextual | ECLS-K |
| SNAP/food stamps | Family received SNAP in last 12 months (binary) | Elementary | Contextual | ECLS-K |
| Household income-to-poverty ratio | Approximate household income-to-poverty ratio (continuous) | Elementary | Contextual | ECLS-K |
| No health insurance | Child does not have health insurance coverage. ; 0 = child has health insurance coverage, 1 = child does not have health insurance coverage (binary) | Elementary | Contextual | ECLS-K |
| Parental support | Scale of responses to six items regarding parental support: watching child during errand, getting a ride to bring child to doctor, checking on child when sick, talking over child's problems at school, emergency cash, and giving advice. Original scale:1 = least, 3 = most. Standardized in model (continuous) | Elementary | Contextual | ECLS-K |
| Neighborhood safety | Indicator of how safe it is for children to play outside during the day in neighborhood (proportion; 0 = not at all safe or somewhat safe, 1 = very safe) | Elementary | Contextual | ECLS-K |
| Neighborhood issues | Are the following issues present in neighborhood: litter/glass, selling/using drugs or alcohol in public, burglary or robbery, violent crimes? Reversed in model so 0 = no issues, -1 = at least one issue (binary) | Elementary | Contextual | ECLS-K |
| Math score | Standardized math proficiency test t-score (continuous) | Middle Childhood | Main Model | ECLS-K |
| Reading score | Standardized reading proficiency test t- score (continuous) | Middle Childhood | Main Model | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Internalizing behavior | Standardized internalizing behaviors score SRS (continuous; higher is less internalizing behavior) | Middle Childhood | Main Model | ECLS-K |
| Externalizing behavior | Standardized externalizing behaviors score SRS score (continuous; higher is less externalizing behavior) | Middle Childhood | Main Model | ECLS-K |
| Peer relationships | Standardized self-described competence in peer relationships (continuous) | Middle Childhood | Main Model | ECLS-K |
| Self-control | Standardized self-control score SRS (continuous) | Middle Childhood | Main Model | ECLS-K |
| Health | Standardized scale of overall health status of child. Original scale:1 = excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Middle Childhood | Main Model | ECLS-K |
| Hearing and seeing problems | Child diagnosed by parent or doctor with difficulty hearing and/or seeing (binary; 0 = does not have health problems, 1 = has health problem) | Middle Childhood | Contextual | ECLS-K |
| Child obesity | Child measured in 95th percentile based on CDC growth chart and is obese (binary) | Middle Childhood | Contextual | ECLS-K |
| Teacher turnover | Teacher turnover is a problem at this school, as reported by school administrator (binary) | Middle Childhood | Contextual | ECLS-K |
| Biological father in the home | Presence of biological father in the home (binary) | Middle Childhood | Contextual | ECLS-K |
| Parent school involvement | Index of responses to items related to parent-school involvement: open houses, PTA meetings, parent-teacher conferences, school events, volunteers, fundraising (continuous, 0-1) | Middle Childhood | Contextual | ECLS-K |
| Out-of-school activities | Child has ever participated in one or more out-of-school activities: dance, athletics, clubs/recreation programs, music lessons, art classes, performing arts (binary) | Middle Childhood | Contextual | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Positive stimulation | Index of responses to items regarding positive stimulation activities: (1) number of children's books in home; (2) outings to library, museum, concert, zoo, and/or sporting event in the past month; (3) positive stimulation activities done with child in a typical week (stories, singing, arts/crafts, involved in chores, games/puzzles, talk about nature/science, building, sport/exercise, practice reading/writing/numbers, read to child (continuous, 0–3) | Middle childhood | Contextual | ECLS-K |
| Parental support | Scale of responses to six items regarding parental support: watching child during errand, getting a ride to bring child to doctor, checking on child when sick, talking over child's problems at school, emergency cash, and giving advice. Original scale:1 = least, 3 = most. Standardized in model (continuous) | Middle Childhood | Contextual | ECLS-K |
| Routines | Index of responses to three items regarding household rules and routines: (1) rules around television (which programs, how early/late, hours during weekdays); (2) regular bedtime; (3) number of meals per week family eats breakfast and dinner together (continuous, 0–3) | Middle Childhood | Contextual | ECLS-K |
| SNAP/food stamps | Family received SNAP in last 12 months (binary) | Middle Childhood | Contextual | ECLS-K |
| Household income-to-poverty ratio | Approximate household income-to-poverty ratio (continuous) | Middle Childhood | Contextual | ECLS-K |
| No health insurance | Child does not have health insurance coverage. ;<br>0 = child has health insurance coverage, 1 = child does not have health insurance coverage(binary) | Middle Childhood | Contextual | ECLS-K |
| Family food insecurity status | Family is food insecure with or without hunger (binary; 0 = family is food secure, 1 = family is food insecure with or without hunger (moderate or severe)) | Middle Childhood | Contextual | ECLS-K |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Family home ownership | Family owns home (binary) | Middle Childhood | Contextual | ECLS-K |
| Neighborhood safety | Indicator of how safe it is for children to play outside during the day in neighborhood (binary; 0 = not at all safe or somewhat safe, 1 = very safe) | Middle Childhood | Contextual | ECLS-K |
| Neighborhood issues | Are the following issues present in neighborhood: litter/glass, selling/using drugs or alcohol in public, burglary or robbery, violent crimes? Reversed in model so 0 = no issues, -1 = at least one issue (binary) | Middle Childhood | Contextual | ECLS-K |
| Negative discipline | Does either parent use any negative disciplinary methods, such as yelling, making fun of, or hitting back? (binary) | Middle Childhood | Contextual | ECLS-K |
| PIAT math score | Standardized score (continuous) | Early Adolescence | Main Model | NLSY |
| ASVAB score | Standardized score (continuous) | Early Adolescence | Main Model | NLSY |
| Delinquency index | Standardized delinquency score index (continuous; higher indicates less delinquent behavior) | Early Adolescence | Main Model | NLSY |
| Positive peer behavior | Standardized scale of positive peer behaviors, including attend church/religious services on regular basis; participate in organized sports, clubs, or school activities; plan to go to college; volunteer work (continuous) | Early Adolescence | Main Model | NLSY |
| Negative peer behavior | Standardized scale of negative peer behaviors, including smoke cigarettes; drunk at least once a month; belong to a gang that does illegal activities; used marijuana, inhalants, or other drugs; cut classes or skip school. Reversed in model so higher value indicates less negative peer behavior (continuous) | Early Adolescence | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Mental health | Standardized scale of responses to items concerning how often the respondent felt certain ways during the previous month (these questions are a five-item short version of the Mental Health Inventory [MHI-5]). Reversed in model so that higher score is better (continuous) | Early Adolescence | Main Model | NLSY |
| Arrested by early adolescence | Indicates whether the respondent has been arrested by early adolescence (binary; 0 = not arrested by life stage, -1 = arrested by life stage) | Early Adolescence | Main Model | NLSY |
| Health | Standardized scale of overall health status Original scale: 1=excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Early Adolescence | Main Model | NLSY |
| Absent from school (number of days) | Number of days absent from school. Reversed in model so that days absent appears as negative (continuous) | Early Adolescence | Main Model | NLSY |
| Suspended from school for 6+ days | Respondent has been suspended from school for more than 6 days. Reversed in model so 0 = not suspended for more than 6 days, -1 = suspended for more than 6 days (binary) | Early Adolescence | Main Model | NLSY |
| Family net worth | Family's net worth at start of survey, adjusted for inflation (continuous) | Early Adolescence | Contextual | NLSY |
| Authoritative parent | Measure of whether either parent has an authoritative parenting style (binary; 0 = neither parent is authoritative, 1 = at least one parent is authoritative) | Early Adolescence | Contextual | NLSY |
| Father in household | Presence of biological father in the home (binary) | Early Adolescence | Contextual | NLSY |
| Gangs in school or neighborhood | Measure of whether there are any gangs in respondent's school or neighborhood . Reversed in model so 0 = no gangs, -1 = there are gangs (binary) | Early Adolescence | Contextual | NLSY |
| Received high school diploma | Received high school diploma by life stage (binary) | Adolescence | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| GPA | Standardized GPA from transcript in spring semester of data year. For most respondents with available data, this will be their senior year of high school (continuous) | Adolescence | Main Model | NLSY |
| Delinquency index | Standardized delinquency score index Reversed in model so higher indicates less delinquent behavior (continuous) | Adolescence | Main Model | NLSY |
| Asks mother and/or father for advice | Standardized scale of responses to items of how often respondent asks mother or father for advice or help on education, training decisions, job decisions, or relationships (continuous) | Adolescence | Main Model | NLSY |
| Had a child by adolescence | Indicates whether the respondent had a child by adolescence. Reversed in model so; 0 = has not had a child by life stage, -1 = has had a child by the life stage (binary) | Adolescence | Main Model | NLSY |
| Mental health | Standardized scale of responses to items concerning how often the respondent felt certain ways during the previous month (these questions are a five-item short version of the Mental Health Inventory [MHI-5]). Reversed in model so that higher score is better (continuous) | Adolescence | Main Model | NLSY |
| Health | Standardized scale of overall health status. Original scale: 1=excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Adolescence | Main Model | NLSY |
| Suspended from school for 6+ days | Respondent has been suspended from school for more than 6 days in data year. For most respondents with available data, this will be their senior year of high school. Reversed in model so 0 = not suspended for more than 6 days, -1 = suspended for more than 6 days (binary) | Adolescence | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Convicted of or plead guilty to crime | Respondent was convicted of or plead guilty to a crime by life stage. Reversed in model so 0 = not convicted/guilty, -1 = convicted/guilty (binary) | Adolescence | Main Model | NLSY |
| Authoritative parent | Measure of whether either parent has an authoritative parenting style (binary; 0 =neither parent is authoritative, 1 = at least one parent is authoritative) | Adolescence | Contextual | NLSY |
| Father in household | Presence of biological father in the home (binary) | Adolescence | Contextual | NLSY |
| Gangs in school or neighborhood | Measure of whether there are any gangs in respondent's school or neighborhood. Reversed in model so 0 = no gangs, -1 = there are gangs (binary) | Adolescence | Contextual | NLSY |
| Victim of a violent crime | Respondent has been a victim of a violent crime in the last five years. Reversed in model so; 0 = not victim of a crime, -1 = victim of a crime (binary) | Adolescence | Contextual | NLSY |
| Lives in rural area | Respondent lives in rural area (binary) | Adolescence | Contextual | NLSY |
| Income-to-poverty ratio | Ratio of household-income-to poverty threshold (continuous) | Transition to Adulthood | Main Model | NLSY |
| Drank before work or school | Respondent reported drinking alcohol before or during work or school at least once in the last month. Reversed in model so; 0 = did not drink before work or school, -1 = drank before work or school (binary) | Transition to Adulthood | Main Model | NLSY |
| Receiving income from job | Respondent receives income from a job (binary) | Transition to Adulthood | Main Model | NLSY |
| Not low income with a child | Measure of whether respondent has a child and is not living below the 200% of the FPL (binary; 0 = had a child while below twice the poverty level, 1 = either had a child above twice the poverty level or did not have a child) | Transition to Adulthood | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Mental health | Standardized scale of responses to items concerning how often the respondent felt certain ways during the previous month (these questions are a five-item short version of the Mental Health Inventory [MHI-5]). Reversed in model so that higher score is better (continuous) | Transition to Adulthood | Main Model | NLSY |
| Health | Standardized scale of overall health status Original scale: 1 = excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (binary) | Transition to Adulthood | Main Model | NLSY |
| Convicted of or plead guilty to crime | Respondent was convicted of or plead guilty to a crime by life stage. Reversed in model so; 0 = not convicted/guilty, -1 = convicted/guilty (binary) | Transition to Adulthood | Main Model | NLSY |
| Received high school diploma | Received high school diploma by life stage (binary) | Transition to Adulthood | Main Model | NLSY |
| Received associate's degree | Received associate's degree by life stage (binary) | Transition to Adulthood | Main Model | NLSY |
| Received bachelor's degree | Received bachelor's degree by life stage (binary) | Transition to Adulthood | Main Model | NLSY |
| Received 30 credits or more higher education, but no degree | Indicates whether respondent had at least 30 credits of higher education but did not receive a degree (binary) | Transition to Adulthood | Main Model | NLSY |
| Completed training or certificate program | Indicates whether respondent has completed a training or certificate program (proportion) | Transition to Adulthood | Main Model | NLSY |
| Inflation-adjusted income | Respondent's total income from wages or salary, adjusted for inflated (continuous) | Transition to Adulthood | Main Model | NLSY |
| Lives in rural area | Respondent lives in rural area (binary) | Transition to Adulthood | Contextual | NLSY |
| Limited in amount or kind of work | Respondent has condition that limits the amount or kind of work her or she can do. Reversed in model so 0 = not limited; -1 = limited (binary) | Transition to Adulthood | Contextual | NLSY |
| Income-to-poverty ratio | Ratio of household-income-to poverty threshold (continuous) | Adulthood | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Drank before work or school | Respondent reported drinking alcohol before or during work or school at least once in the last month. Reversed in model so 0 = did not drink before work or school, -1 = drank before work or school (binary) | Adulthood | Main Model | NLSY |
| Receiving income from job | Respondent receives income from a job (binary) | Adulthood | Main Model | NLSY |
| Not low income with a child | Measure of whether respondent has a child and is not living below the 200% of the FPL (binary; 0 = had a child while below twice the poverty level, 1 = either had a child above twice the poverty level or did not have a child) | Adulthood | Main Model | NLSY |
| Mental health | Standardized scale of responses to items concerning how often the respondent felt certain ways during the previous month (these questions are a five-item short version of the Mental Health Inventory [MHI-5]). Reversed in model so that higher score is better (continuous) | Adulthood | Main Model | NLSY |
| Health | Standardized scale of overall health status. Original scale: 1=excellent, 5 = poor. Standardized and reversed in model so that a higher number means better health (continuous) | Adulthood | Main Model | NLSY |
| Convicted of or plead guilty to crime | Respondent was convicted of or plead guilty to a crime by life stage. Reversed in model so 0 = not convicted/guilty, -1 = convicted/guilty (binary) | Adulthood | Main Model | NLSY |
| Received associate's degree | Received associate's degree by life stage (binary) | Adulthood | Main Model | NLSY |
| Received bachelor's degree | Received bachelor's degree by life stage (binary) | Adulthood | Main Model | NLSY |
| Received 30 credits or more higher education, but no degree | Indicates whether respondent had at least 30 credits of higher education but did not receive a degree (binary) | Adulthood | Main Model | NLSY |
| Completed training or certificate program | Indicates whether respondent has completed a training or certificate program (binary) | Adulthood | Main Model | NLSY |

| | Description | Life stage | Variable type | Source of variable |
|---|---|---|---|---|
| Inflation-adjusted income | Respondent's total income from wages or salary, adjusted for inflated (continuous) | Adulthood | Main Model | NLSY |
| Lives in rural area | Respondent lives in rural area (binary) | Adulthood | Contextual | NLSY |
| Limited in amount or kind of work | Respondent has condition that limits the amount or kind of work her or she can do. Reversed in model so 0 = not limited; -1 = limited (binary) | Adulthood | Contextual | NLSY |
| Lifetime earnings | Continuous measure of lifetime income generated using coefficients from the Urban Institute's DYNASIM[10] model; lifetime earnings are based on education, health, and earnings at age 30 (continuous) | Estimated for age 65 | Main Model | DYNASIM [a] |

**Notes:** NLSY = National Longitudinal Survey of Youth; ECLS-K =Early Childhood Longitudinal Program, Kindergarten Class of 1998–99.

TABLE A.5

**Early Childhood Coefficient Data Dictionary**

| | | | |
|---|---|---|---|
| Math score | Math proficiency scale score (standardized), measured at 5 years | Preschool | ECLS-B |
| Reading score | Reading proficiency scale score (standardized), measured at 5 years | Preschool | ECLS-B |
| Health | Overall health status of child (standardized), measured at 5 years | Preschool | ECLS-B |

**Note:** ECLS-B = Early Childhood Longitudinal Study, Birth Cohort.

# Notes

[1] Richard V. Reeves and Kimberly Howard, "The Marriage Effect: Money or Parenting?" *Social Mobility Memos* (blog), September 4, 2014, http://www.brookings.edu/blogs/social-mobility-memos/posts/2014/09/04-marriage-social-mobility-parenting-income-reeves.

[2] The Urban Institute's Dynamic Simulation of Income Model (DYNASIM) projects the size and characteristics—such as financial, health, and disability status—of the US population for the next 75 years (Urban Institute 2015).

[3] For information on and data from the ECLS-K, see "Kindergarten Class of 1998–99 (ECLS-K)," National Center for Education Statistics, accessed January 6, 2021, https://nces.ed.gov/ecls/kindergarten.asp

[4] "Observations" refers to respondents in either the NLSY97 or ECLS-K.

[5] We rescaled weights for the NLSY97 and ECLS-K separately, scaling them to a common total weight out of 10,000,000 for each dataset. Weights are rounded up to the nearest integer.

[6] Early childhood variables are included only as coefficients estimated from the ECLS-B data, and there is no individual-level data on these variables in the underlying dataset.

[7] Adult outcomes in the SGM are measured at age 30. Adult outcomes from the CPS, NHIS, and BRFSS reflect age-adjusted percentages among US adults age 18 and over.

[8] For information on and data from the ECLS-B, see "Birth Cohort (ECLS-B)," National Center for Education Statistics, accessed January 6, 2021, https://nces.ed.gov/ecls/birth.asp

[9] For information on and data from the ECLS-B, see "Birth cohort (ECLS-B)," National Center for Education Statistics, accessed January 6, 2021, https://nces.ed.gov/ecls/birth.asp

[10] To generate lifetime earnings in our model, we use a regression equation using data from the DYNASIM model (see endnote two for more information on DYNASIM). The dependent variable in this model is lifetime earnings, discounted at a rate of 2.3 percent. We regress discounted lifetime earnings on health at age 30, binary variables for race/ethnicity, binary variables for educational attainment at age 30, yearly earnings at age 30, and interactions between earnings at age 30 and educational attainment. We run the regression separately for men and for women. From these regressions, we get coefficients which we apply to the data in our model in order to estimate lifetime earnings

# References


Bronfenbrenner, Urie. 1979. *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Becker, Gary S. 1975. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Second Edition*. New York: National Bureau of Economic Research and Columbia University Press.

Card, David. 2001. "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems." *Econometrica* 69 (5): 1127–1600.

Carneiro, Pedro, James Heckman, and Edward Vytlacil. 2011. "Estimating Marginal Returns to Education." *American Economic Review* 101 (6): 2754–2781.

Chetty, Raj and Nathaniel Hendren. 2018. "The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates." *The Quarterly Journal of Economics* 133 (3): 1163–1228.

Cunha, Flavio and James Heckman. 2009. "Investing in Our Young People." *Rivista Internazionale di Scienze Sociali* 117 (3): 387–418.

Currie, Janet, and Duncan Thomas. 1995. "Does Head Start Make a Difference?" *American Economic Review* 85 (3): 341–364.

Duckworth, Angela, and Martin E. P. Seligman. 2005. "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents." *Psychological Science* 16 (12): 939–944.

Elder, Glen H. Jr. 1998. "The Life Course as Developmental Theory." *Child Development* 69 (1): 1–12.

Favreault, Melissa M., Karen E. Smith, and Richard W. Johnson. 2015. *The Dynamic Simulation of Income Model (DYNASIM): An Overview.* Washington, DC: Urban Institute.

Fontenot, Kayla, Jessica Semega, and Melissa Kollar. 2018. *Income and Poverty in the United States: 2017*. P60-263. Washington, DC: US Census Bureau.

Garcia, Jorge, and James Heckman. 2014. "Ability, Character, and Social Mobility." Unpublished paper. University of Chicago. Department of Economics.

Goldin, Claudia, and Lawrence Katz. 2008. *The Race between Education and Technology*. Cambridge, MA: Belknap Press.

Glover, Vivette. 2011. "Annual Research Review: Prenatal Stress and the Origins of Psychopathology—An Evolutionary Perspective." *Journal of Child Psychology and Psychiatry* 52 (4): 356–67.

Heckman, James J. 2000. "Invest in the Very Young." Chicago: Ounce of Prevention Fund.

Heckman, James J., and Tim Kautz. 2012. "Hard Evidence on Soft Skills." *Labour Economics* 19 (4): 451–64.

Heckman, James J., and Yona Rubinstein. 2001. The Importance of Noncognitive Skills: Lessons from the GED Testing Program. *American Economic Review* 91 (2): 145–149.

Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the High/Scope Perry Preschool Program." *Quantitative Economics, Econometric Society* 1 (1): 1–46.

Heckman, James J., Jora Stixrud., and Segio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24: 411–482.

Henderson, Daniel, Solomon Polachek, and Le Wang. 2011. "Heterogeneity in Schooling Rates of Return." *Economics of Education Review* 30 (6): 1202–214.

Kena, Grace, Susan Aud, Frank Johnson, Xiaolei Wang, Jijun Zhang, Amy Rathbun, Sidney Wilkinson-Flicker, and Paul Kristapovich. 2014. *The Condition of Education 2014*. Washington, DC: US Department of Education, National Center for Education Statistics.

Matthews, T. J., and Brady E. Hamilton. 2016. "Mean Age of Mothers Is on the Rise: United States, 2000–2014." Data brief 232. Hyattsville, MD: National Center for Health Statistics.

Mincer, Jacob. 1981. "Human Capital and Economic Growth." Working Paper 803. Cambridge, MA: National Bureau of Economic Research.

Moore, Kristin Anderson. 1997. "Criteria for Indicators of Child Well-Being." In *Indicators of Children's Well-Being* edited by Robert M. Hauser, Brett V. Brown, and W.R. Prosser, 36–44. New York: Russell Sage.

———. 2020. "Developing an Indicator System to Measure Child Well-Being: Lessons Learned over Time." *Child Indicators Research* 13: 729–739.

Moore, Kristin Anderson, Hannah Lantos, Rebecca Jones, Ann Schindler, Jonathan Belford, and Vanessa Sacks. 2017. *Making the Grade: A Progress Report and Next Steps for Integrated Student Supports*. Bethesda, MD: Child Trends.

Nelson, Charles A. III, and Katherine Magnusun. 2011. "Lessons from Neuroscience Research for Understanding Causal Links between Family and Neighborhood Characteristics and Educational Outcomes." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances* edited by Greg J. Duncan and Richard J. Murnane, 27–46. New York: Russell Sage Foundation.

Patrinos, Harry Anthony. "Estimating the Return to Schooling Using the Mincer Equation." Bonn, Germany: IZA World of Labor.

Reynolds, Arthur J., Judy A. Temple, Suh-Ruu Ou, Irma A. Arteaga, and Barry A. B. White. 2011. "School-Based Early Childhood Education and Age-28 Well-Being: Effects by Timing, Dosage, and Subgroups." *Science* 333 (6040): 360–64.

Ross, Martha, Kristin A. Moore, Kelly Murphy, Nicole Bateman, Alex DeMand, and Vanessa Sacks. 2018. "Pathways to High-Quality Jobs for Young Adults." Washington, DC: Brookings Institution.

Sawhill, Isabel. 2014. *Generation Unbound: Drifting into Sex and Parenthood without Marriage*. Washington, DC: Brookings Institution Press.

Sawhill, Isabel, and Quentin Karpilow. 2014. "How Much Could We Improve Children's Life Chances by Intervening Early and Often?" Washington, DC: Brookings Institution.

Sawhill, Isabel, and Stephanie Owen. 2013. "Should Everyone Go to College?" Washington, DC: Brookings Institution.

Sen, Amartya. 1992. *Inequality Reexamined*. Massachusetts: Harvard University Press.

Shonkoff, Jack P., and Deborah A. Phillips, eds. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development.* Washington, DC: National Academy Press.

Urban Institute. 2015. "DYNASIM: Projecting Older Americans' Future Well-Being." Washington, DC: Urban Institute.

Werner, Kevin, Kristin Blagg, Gregory Acs, Steven Martin, Alison McClay, Kristin Anderson Moore, Gabriel Piña, and Vanessa Sacks. "Social Genome Model 2.0: Technical Documentation and User's Guide." Washington, DC: Urban Institute.

# About the Authors

**Kevin Werner** is a research associate in the Income and Benefits Policy Center. He is part of the team that maintains and develops the TRIM3 model, a microsimulation model that simulates major government tax and transfer programs and allows researchers to see the effects of various policy changes. He also works with the Urban Institute's ATTIS model. He has dual degrees in economics and political science from American University and a master's degree in applied economics from Georgetown University.

**Kristin Blagg** is a senior research associate in the Center on Education Data and Policy at the Urban Institute. Her research focuses on K–12 and postsecondary education. Blagg has conducted studies on student transportation and school choice, student loans, and the role of information in higher education. In addition to her work at Urban, she is pursuing a PhD in public policy and public administration at the George Washington University. Blagg holds a BA in government from Harvard University, an MSEd from Hunter College, and an MPP from Georgetown University.

**Gregory Acs** is vice president for income and benefits policy at the Urban Institute, where his research focuses on social insurance, social welfare, and the compensation of workers. Acs has studied the low-wage labor market, changes in welfare policies and how they have affected welfare caseloads and the well-being of low-income families, and how state and federal policies affect the incentives families face as they move from welfare to work. Acs holds a PhD in economics and social work from the University of Michigan.

**Steven Martin** is a senior research associate at the Urban Institute with a specialization in family demography. He also does work on numerous topics related to quantitative analysis of policy outcomes such as analyses of event histories of interdependent events, time use, quality of life measures, attitudinal measures, and a variety of other work with survey, census, and vital statistics data. Dr. Martin has also published on data quality, including the analysis of predictors of response error for different types of respondent recall under different survey formats. Martin holds a PhD in sociology from the University of Wisconsin.

**Alison McClay** is a senior research analyst within the youth development program area of Child Trends. She earned her Master of Public Health in Maternal and Child Health from Gillings School of Global Public Health at the University of North Carolina at Chapel Hill. During her time at Chapel Hill,

Alison supported a federally contracted survey study as a survey research assistant with the H.W. Odum Institute for Research in Social Science. She also coordinated program evaluation activities for Community-Academic Resources for Engaged Scholarship projects as a graduate research assistant at the North Carolina Translational and Clinical Sciences Institute. For her practicum, Alison worked with the Vermont Department of Health as a Title V Maternal and Child Health intern investigating the relationship between health care access and utilization among women working in agriculture.

**Kristin Anderson Moore** is an internationally recognized social psychologist with more than 40 years of experience monitoring, studying, and evaluating child and family well-being. Dr. Moore is trained as a survey researcher and has worked on numerous federal surveys, as well as surveys designed for evaluation studies. Her current work at Child Trends includes an evaluation of youthCONNECT in Prince George's County in Maryland, development of the Social Genome Model, a study of positive youth development in five Generation Work communities, and development of a Healthy and Ready to Learn measure for children ages 3–5 in the National Survey of Children's Health. She led two studies of integrated student supports and codirected development of El Camino, an intervention to reduce teen pregnancy while enhancing educational engagement. Dr. Moore earned her PhD in psychology from the University of Michigan.

**Gabriel Piña** is a research scientist in youth development at Child Trends. He works on quantitative analyses to research youth development and early childhood education issues. Gabriel is one of the primary analysts on the Social Genome Model, a microsimulation model that draws on available research and evaluation studies to forecast how early childhood interventions can affect critical adolescent and adult outcomes like level of education, income, unemployment, and poverty. Prior to joining Child Trends, Gabriel's research focused on examining the impact of homeless prevention programs on child and youth residential instability, and conducting benefit-cost analyses of early childhood education programs in Minnesota. He earned his PhD in public affairs from Indiana University.

**Vanessa Sacks** works in the Youth Development research area at Child Trends. Vanessa's work spans a wide range of quantitative analyses and youth development issues. She has conducted numerous studies, from quasi-experimental evaluations of youth program outcomes, to analyses of state policy using large national data sets, to performance management projects. Prior to Child Trends, Vanessa worked in nonprofit fundraising for many years and has extensive project management and communications experience. Her primary research interests are around the social and health barriers that low-income youth face in their path to economic self-sufficiency and evaluating intervention programs aimed at this population.

## Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.




500 L'Enfant Plaza SW
Washington, DC 20024

*www.urban.org*