Guide to the Brookings Social Genome Model

Scott Winship and Stephanie Owen[1]
January 16, 2013

Abstract

Policymakers often lack reliable information on the likely impacts of policies intended to promote upward mobility. Few studies attempt or are able to estimate the effects of a policy intervention in childhood on outcomes decades later. Even randomized controlled trials examining a single intervention generally are uninformative as to whether one policy administered to the same target population is more or less effective than another, whether the timing of interventions matters, or whether multiple interventions—concurrent or sequential—have notably bigger impacts than single ones. The Brookings Social Genome Model is a new dynamic microsimulation model developed to answer these questions. The paper describes the SGM and how it was developed, explains how policy simulations are conducted using the model, and discusses the challenges facing any effort to model mobility processes and simulate long-term effects of early interventions. These difficulties include both data issues and methodological challenges to estimating valid effects.

[1] Winship is a fellow, Owen a senior research assistant at Brookings. We thank our colleagues (present and past) Isabel Sawhill, Alex Gold, Jeff Diebold, Adam Thomas, Julia Isaacs, Kerry Grannis, Quentin Karpilow, Kim Howard, and Daniel Moskowitz for their contributions to this research. We also thank our advisers—J. Lawrence Aber and Katherine Magnuson especially, along with Gordon Berlin, Robert Haveman, Sara McLanahan, Richard Murnane, Karen Pittman, Michael Wald, and Jane Waldfogel for very helpful comments on an earlier version of this paper. Finally, we thank Martin Holmer, Karen Smith, and especially Cathal O'Donoghue for technical advice on simulation questions, and Lynn Karoly for her work helping to validate the model.

Contents

Introduction

The promise of upward mobility is a central tenet of the American Dream, one of our core civic values. Generation after generation, Americans have been more likely than not to end up better off financially than their parents were. That has been the experience of four in five of today's middle-age adults. At the same time, it is no less true today than in generations past that Americans' opportunities remain stubbornly linked to the incomes of their parents. Roughly four in ten of today's middle-age adults who were raised by the poorest fifth of families remain in the poorest fifth themselves. The same share of today's adults raised by the richest fifth of families is in the richest fifth themselves.[2] A society in which poor children can anticipate being less poor in adulthood but must resign themselves early on to the likelihood they will still occupy the bottom rungs should satisfy no adherents to the American Dream.

The question of what promotes and impedes economic mobility is dauntingly complex. Policymakers seeking to broaden upward mobility face great challenges formulating effective solutions. Increasingly, they can draw from a range of high-quality randomized controlled trials that look at this or that intervention aimed at specific groups of children. Yet this evidence is limited by the scarcity of long-term studies, by inconsistencies between studies examining different policies and programs, and by a dearth of tools that can forecast the likely impacts of untested policies.

The Social Genome Model (SGM) is a microsimulation model of the life cycle that tracks the academic, social, and economic experiences of individuals from birth through middle age in order to identify the most important paths to upward mobility. Equally important, it facilitates simulations to estimate the likely medium- and long-term effects of policy interventions to promote mobility. The model divides the years from birth to forty into five stages. At each point where stages meet, we consider a range of outcomes chosen for their established links to subsequent outcomes and to reflect broadly-shared norms about what success entails at different ages.

The SGM will fill important gaps in the field of program evaluation. For one, it will allow for credible estimates of long-term effects of programs and policies that intervene in early childhood without the necessity of waiting forty years to assess results. Unlike the set of randomized controlled trials that currently exists, it will allow for apples-to-apples comparisons of interventions applied in a given life stage. The SGM will also facilitate evaluation of intervening concurrently in multiple ways or successively at multiple stages. It will allow for estimates of the relative effectiveness of intervening earlier or later. Finally, it will facilitate decision-making around interventions that have yet to be tried.

Model Specification

The theory behind our model is quite simple and draws on a large literature on human capital formation and its effects on later earnings and income. We focus on the development of both cognitive and noncognitive skills in early and later childhood, in the same spirit as James Heckman and others, after controlling for a child's circumstances at birth[3].

[2] Pew Economic Mobility Project (2012). See also Isaacs, Sawhill, and Haskins (2008).

[3] See, among others, Heckman and Rubinstein (2001), Heckman, Stixrud, and Urzua (2006), Duckworth and Seligman (2005) and Shonkoff and Phillips (2000).

The human capital formation process is modeled by measuring these cognitive and noncognitive skills at successive life stages from early childhood (the preschool period) through middle childhood and adolescence. We look at the results of that process at ages 5, 11, and 19 (or as close to those ages as possible). We then look at how success in adolescence translates into success in early and middle adulthood – specifically at ages 29 and 40 (or as close to these ages as possible). Because there is a break in our longitudinal data at the end of adolescence, this last set of transitions poses some special challenges which we return to in the section "Projecting Adult Outcomes."

Because we are interested in the process of human capital formation and not just the end result, we use a set of structural equations, one for each life stage, so that we can see the direct and indirect effects of earlier success on later success. A substantial portion of our work has also been devoted to, and will continue to be devoted to, a more detailed look at the determinants of success *within* each life stage. That work will inform the larger goal of estimating the effects of particular interventions on later life success.

Because the model is a life cycle model, its specification relies heavily on a temporal element. Prior outcomes are normally assumed to have a causal effect on later outcomes. Nonetheless, the SGM like all models relies on a number of assumptions, and as described in the section "Challenges," we still have to worry about distinguishing correlation from causation. However, having a model of the life course can serve as the starting point for sensitivity analyses, and the model may be improved over time. It is our hope that the development of the model will help focus researchers' efforts to assess what we do and do not yet know about the processes behind social mobility.

The SGM Dataset

The SGM is formed using two data sets from the Bureau of Labor Statistics' National Longitudinal Surveys. Our primary data set is the "Children of the NLSY79" (CNLSY), representing children born mainly in the 1980s and 90s. The CNLSY is the source for our data on birth circumstances, early and middle childhood, and adolescence. No respondent in the CNLSY is yet old enough to track through adulthood, and so we impute adult values using a second sample from an earlier generation, the "National Longitudinal Survey of Youth 1979" (NLSY79)[4]. To do the imputation, we use regression analysis of the NLSY79 to relate child background characteristics and adolescent outcomes to adult outcomes, and then apply these coefficients to the same measures in the CNLSY sample to estimate adult outcomes.

[4] In fact, the CNLSY children were the progeny of NLSY79 women. The NLSY79 began with a nationally-representative sample of over 12,000 men and women, aged 14 to 22 in 1979 (born between 1957 and 1964). As women in the NLSY79 have given birth to children—11,504 as of 2009, by which time the childbearing of the NLSY79 women was essentially complete—detailed information has also been collected on them in the CNLSY.

The result is a longitudinal dataset in which these synthetic individuals pass through five life stages from birth to adulthood: early childhood (birth through age five), middle childhood (age six through age eleven), adolescence (age twelve through age nineteen), transition to adulthood (age twenty through age twenty-nine), and adulthood (age thirty through age forty). Our final dataset includes 5,783 children from the CNLSY, born between 1971 and 2009, rather than the 11,504 included in the original data.[5] See the "Data Challenges" section, below, for additional detail on the creation of our dataset.

The Model

The Variables and Success Measures

In its current state, the SGM includes a range of outcomes from six different stages. Table 1 summarizes them[6]. Descriptive statistics for all of the variables are in Table 2 and Table 3. Using a

[5] The NLSY79 included a cross-sectional sample of civilian men and women as well as additional samples of African Americans, Hispanics, and poor youth who were neither Hispanic nor black, plus a military sample. Most of the military sample was dropped after 1984, and the entire supplemental sample of poor non-Hispanic non-blacks was dropped after 1990. Using the other supplemental samples makes weighting the data essential, and for reasons we discuss below, we were uncomfortable using the weights provided with the CNLSY and NLSY79 data. Because of these issues, we chose to use only the cross-sectional samples of CNLSY children and NLSY79 adults. Note that because the CNLSY children were born to mothers who were living in the U.S. in 1978, using the survey means that we necessarily exclude children who immigrated here after 1978, as well as children born to mothers who immigrated after that year. Our data and model, then, are best viewed as applying to the entire set of children born to women living in the U.S.

[6] Some details worth noting: Our parental marital status indicator groups cohabiting but unmarried couples with single parents. All of our early and middle childhood measures are first standardized on children of the same age who have non-missing raw scores. We then aggregate children from adjacent age groups (e.g., five-, six-, seven-, and eight-year-olds) and impute standardized scores to those who are not observed at any of the ages. In adolescence, our high school graduation variable indicates whether a person received a traditional high school diploma; we do not count holding a GED as graduating from high school. This is consistent with research showing that GED holders do worse than traditional graduates and often no better than dropouts in the long run (Tyler 2003). Both young men and women report whether or not they became a parent by age 19, but half as many men report having become parents. Several of the adolescent variables are used mainly for purposes of linking the CNLSY and NLSY79. We try to define the CNLSY and NLSY79 linking variables as similarly as possible given the differences between the two data sets. Grade point averages in the last year of high school are reported by CNLSY respondents as a letter grade (A+, A, A-, etc.), and while one might worry that they exaggerated in their responses, a quick check against the 1997 panel of the National Longitudinal Survey of Youth, which included children born in the early 1980s and which includes transcript-derived GPAs, found comparable results. In the NLSY79, GPA in the last year of school is computed directly from high school transcript information in the data. The adult GPA distribution is smoother as a result. It also has a lower mean, which we interpret as mainly reflecting grade inflation over time (given the corroborating evidence from the 1997 NLSY). The adolescent test scores in the CNLSY are from the Peabody Individual Achievement Test (PIAT) reading recognition and math subscales, administered around ages 13-14; those in the NLSY79 are from the Armed Forces Vocational Aptitude Battery (ASVAB) word knowledge and arithmetic reasoning subscales, administered between ages 15 and 23. All four score distributions are age-adjusted. All of the family income variables we use, which come from different survey years, are measured in constant 2010 dollars, adjusted using the Census Bureau's CPI-U-RS. They include income from a large number of sources, but they exclude income received by cohabiting partners of the NLSY79 respondent or the CNLSY child's mother. We applied a common top code to incomes in all years that was as

subset of our outcomes, we have defined success indicators for each life stage based on outcomes that have been shown to predict future success and that are widely considered to be important from a normative perspective (See Table 4). In early and middle childhood, we require that a child not be too far behind his or her peers academically, behaviorally, and socially. In adolescence, we require that individuals finish high school with a minimum GPA of 2.5 and avoid being convicted or becoming a parent. In early adulthood, we require individuals to be living independently of their parents, and to either have a college degree or an equivalent family income (250% of the federal poverty line, or about $45,000 for a married couple with one child, which is similar to the annual earnings of the typical full-time worker with a college degree at this age).[7] In adulthood, being "middle class by middle age" means having family income at least 300% of the poverty line, or around $68,000 for a married couple with two children.[8] While the thresholds required for success on each continuous subcomponent at each stage are, admittedly, arbitrary, they serve as useful heuristics in the absence of logical breaks within the data or established research findings.

The Model's Structure

Using the dataset we created, discussed in detail, below, in the section on "Projecting Adult Outcomes," SGM predicts the 33 outcomes from early childhood through adulthood listed in Table 1. Through adolescence, it does so using the Circumstances at Birth (CAB) variables in Table 1 plus all outcomes from intervening stages. So, for example, if we were predicting high school graduation, one of the outcomes in adolescence, the regression equation would include all of the CAB variables and all of the outcomes in early childhood (EC) and middle childhood (MC). The equation we estimate for each outcome through adolescence (ADOL) is:

$$Outcome = \beta_0 + \beta_1 CAB + \beta_2 Previous\ Stage\ Outcomes + \varepsilon \qquad \text{Equation 1}$$

where $\beta_1$and β_2 are vectors of coefficients, *CAB* is the set of Circumstances at Birth variables in Table 1, *Previous Stage Outcomes* is the set of outcomes from temporally prior stages, and ε is the error term containing unobserved characteristics.

Beginning with transition to adulthood (TTA) outcomes, however, we must estimate different equations because of our reliance on NLSY79-based imputations for measures in TTA and in adulthood. We are limited to predictor variables that are common to both datasets, which come from the CAB and ADOL stages. For TTA outcomes we estimate:

$$TTA\ Outcome = \beta_0 + \beta_1 CAB^* + \beta_2 ADOL + \varepsilon \qquad \text{Equation 2}$$

where the asterisk following CAB indicates the subset of CAB variables that are available in the NLSY79 and where ADOL is the set of adolescent outcomes.[9] For adulthood income, we estimate:

$$Adult\ Income = \beta_0 + \beta_1 CAB^* + \beta_2 ADOL + \beta_3 TTA + \varepsilon \qquad \text{Equation 3}$$

---

restrictive as that applied in the most restrictive year. We compute income-to-needs ratios by comparing family incomes and family sizes against the poverty guidelines published by the U.S. Department of Health and Human Services.

[7] 2011 poverty threshold for family of 3 with one child is $18,106 (U.S. Census Bureau).

[8] In 2011, poverty threshold for family of 4 with 2 children was $22,811 (U.S. Census Bureau).

[9] The subset of CAB variables in the NLSY79 includes race, gender, maternal age, and maternal education.

where *TTA* is the set of transition-to-adulthood outcomes. Note that EC and MC outcomes cannot directly affect TTA outcomes and adulthood income in these specifications, though they may indirectly affect them through the ADOL variables. The SGM may be shown in a graphically as in Figure 1.

Process for Doing Simulations

In order to simulate the effect of any policy intervention, we use the following procedure:

1. Estimate coefficients for our regression equations
2. Use those coefficients to create a synthetic baseline
3. Adjust one or more variables to reflect the policy intervention
4. Propagate the effects of that intervention through the model using the coefficients estimated in Step 1
5. Calculate the effect of the intervention on later outcomes
6. Calculate the effect on lifetime income

Step 1: Estimating Coefficients

We estimate coefficients on our entire nationally representative samples of children in the CNLSY and adults in the NLSY79[10]. As we discuss below, we conduct substantial imputation of missing values in both surveys, and we include cases with imputed values in these estimation samples. Continuous outcomes (all early and middle childhood outcomes, GPA, and the income measures) are estimated using OLS.[11] To account for the long right tail of income variables, we estimate them in logged forms which are converted back to their original metric when we report the results. Binary outcomes are estimated using a linear probability model.[12]

Step 2: Creating the Synthetic Baseline

Once we have estimated the model, we use the estimated coefficients and the actual values for the baseline characteristics to predict each of the outcomes for every individual in the target population. The target population can be defined either by the limited applicability of an intervention (e.g. children who already attend preschool cannot be affected by an intervention that takes the form of enrolling kids in preschool) or because the effect size we use for a given policy is taken from a rigorous evaluation of a specific population and would require unacceptable assumptions to generalize (e.g. the Nurse Family Partnership home visiting program generally has been available only to poor, first-time mothers).

---

[10] We might prefer to newly estimate the coefficients on simulation-specific target populations each time. However, because our TTA and adulthood income equations must be estimated on NLSY79 data, and only limited pre-adolescent information is available in that data, it is not generally possible to restrict this data to target populations defined with respect to at-birth characteristics or early outcomes.

[11] Continuous measures include all early and middle childhood outcomes, GPA, all income measures, and a number of adolescent variables including math and reading scores, self-esteem, frequency of religious service, and gender role attitudes.

[12] Binary measures include high school graduation, teen birth, conviction, college graduation, marijuana use, other drug use, early sex, suspension, fighting, hitting, damaging property, participation in school clubs, and independence in ADOL and TTA. We confirmed that our results were similar using logistic regression models and chose linear probability models for the greater flexibility they have in the context of structural equation modeling.

For the 15 continuous outcomes in EC, MC, and ADOL, we add the residual terms back to individuals' predicted values, which leaves each person's baseline value the same as their actual value.[13] We do so because we reassign each person the same residual when we implement the intervention later on. Doing so ensures that the only thing that changes between the baseline and policy estimates is the value of the outcome or outcomes that the policy intervention affects, and it leaves the simulated counterfactual as consistent with the actual baseline as possible. It also incorporates into the policy estimates potentially valuable information about individuals' unobserved characteristics.

For the 12 binary outcomes in adolescence, the linear probability models are used to produce predicted probabilities for each individual. These estimates are bound such that no individual may have a predicted probability less than 0 or greater than 1. In order to assign each person a dichotomous value, they are randomly assigned a number between 0 and 1. If their random number is less than their predicted probability, then the outcome is predicted to occur. If their random number is greater than or equal to their predicted probability, then their outcome is predicted not to occur. We retain the random number drawn for each person for the simulated counterfactual, again, in order to keep everything as consistent as possible with the baseline.

For TTA and adulthood outcomes, the creation of baseline values is somewhat different because of the necessity of relying on the NLSY79 to estimate coefficients. To impute TTA outcomes, we use actual CAB values from the CNLSY with the corresponding coefficients estimated from the NLSY79, but we use the *baseline* adolescent values rather than the actual values in the CNLSY data. For continuous adolescent outcomes, the baseline values are exactly the same as the actual values because we add residuals to the predicted values, but for dichotomous adolescent outcomes, the baseline values are those predicted from the procedure described above.[14]

To impute adult income, we again use actual CAB values from the CNLSY and baseline adolescent values, and we also use the baseline TTA values just estimated. All of these values are combined with the coefficients estimated from the NLSY79. Since we do not have actual TTA and adulthood outcomes, we do not have actual residual terms for each individual after estimating continuous baseline outcomes. We instead give everyone a residual that is randomly drawn from a normal distribution with mean zero and with standard deviation taken as the standard error of regression from the applicable NLSY79 equation. As with earlier stages, after predicting dichotomous outcomes using a linear probability model, we take a random draw to determine whether or not to assign individuals a 0 or a 1.

Step 3: The Intervention

To implement a policy intervention or "what-if" scenario, we must first make three important decisions: which metric or metrics are affected, for whom, and by how much. For "what-if" scenarios, this is simply a matter of specifying the change, such as "what if we equalized the middle childhood

[13] GPA is restricted to be between 0 and 4 after prediction.

[14] Those baseline values need not equal the actual values in the CNLSY because our predictions of dichotomous outcomes are imperfect. It might seem preferable to use the actual values here, but doing so would create inconsistencies in the post-intervention run of the model—we might predict, in the post-intervention run, some actual high school graduates, for instance, to be dropouts, which would mean that an intervention could be estimated to worsen outcomes among some youth.

reading scores of poor and non-poor children?" In that case we would just increase every poor child's reading score by the amount of the poor/non-poor reading gap. For a policy intervention, we rely on the best-practice evaluations, preferably randomized controlled trials, of others to generate effect sizes. When determining an effect size, we err on the conservative side or simulate a range of possible effects to avoid a false sense of precision and to account for differences between metrics in our model and the evaluation studies.

We also use the data in the evaluation literature to determine which portion of our model's population should receive the effects of the program, looking at whether the evidence shows heterogeneous effects on particular subgroups. The comprehensive school reform program, Success for All, for example, was implemented in a variety of schools nationwide and showed a high degree of homogeneity of its effects in different schools; on the other hand, a program like Nurse Family Partnership, for which only low-income, first-time mothers are eligible, requires that we narrow our "treatment group" in the model.

After deciding on the target population and the appropriate effect size, we apply the intervention differently depending on whether it affects a continuous or dichotomous variable. If it is a continuous variable, we simply add the effect size to everyone in the target group. For interventions on dichotomous variables, we come up with effect sizes as a percent change from baseline. For example, if some intervention increases high school graduation by 15 percent, we calculate how many extra individuals (N) in our data would need to graduate to increase the rate within the target population by 15 percent, randomly sort the individuals who were in the target group and had not graduated from high school, and then change the top N people from non-graduates to graduates.

Step 4: Propagating the Effects Through the Model

In order to simulate the effect of the changes we make in Step 3 on subsequent life stages, we apply the estimated coefficients from Step 1 to the simulated data, which have now been adjusted according to the effect size of the intervention being evaluated. In doing so, we implicitly assume that the only thing an intervention changes is a person's measured outcomes, and not the relationship between the different outcomes or unmeasured outcomes.

Every outcome prior to the intervention stage is unaffected, as is every outcome in the intervention stage that we did not perturb directly as part of the intervention. We iterate though the subsequent stages and predict outcomes for each stage using earlier outcomes, which have been adjusted by the intervention. This ensures that the effect of the intervention is carried though the entire life course. For example, if we improved middle childhood reading, our post-intervention data through middle childhood would be exactly the same as the pre-intervention baseline (except for middle childhood reading) but our adolescent data would be predicted using the increased reading scores and would reflect that change. To predict the Transition to Adulthood outcomes, we would use the newly-predicted adolescent outcomes that include the effect of the intervention, and adulthood income would be predicted from these new adolescent outcomes as well the newly-predicted Transition to Adulthood outcomes. As noted above, to ensure that our effect size reflects only the impact of the intervention, continuous outcomes are assigned their same residual from Step 2, and dichotomous outcomes are assigned a 0 or 1 based on the same random number from Step 2.

Step 5: Calculating the Impact of the Intervention

When reporting how outcomes have changed based on an intervention which alters one or more earlier outcomes, we compare the pre-intervention simulated outcomes from Step 2 to the post-intervention simulated outcomes from Step 4. For most outcomes, the pre- and post- values are used to calculate a percent change in each outcome as a result of the intervention. If a middle childhood intervention increases the high school graduation rate from 75% to 80%, then the effect size is to increase graduation by (80-75)/75 = 6.7%. For our early and middle childhood outcomes, which are all measured in terms of standard deviations, we simply subtract the pre- value from the post- one.

Next, we assess how the intervention affected general measures of "success" at each life stage. The success measures are dichotomous variables corresponding to the definitions given in Table 4. We estimate success rates using the pre-intervention simulated outcomes for the individual components of success, and we do the same using the post-intervention simulated outcomes.[15]

Step 6: Calculating the Impact on Lifetime Income

Along with the effects on our outcomes and success measures, we also report the effect of our interventions on lifetime income. In order to get a pre-intervention estimate for lifetime family income, we use the means of two data points we know for each individual in our dataset: family income at age 29 and family income at age 40. We calculate the slope between these two ages as:

$$29\text{-}to\text{-}40\,slope = (\overline{Income_{40}} - \overline{Income_{29}})/11 \qquad \text{Equation 4}$$

and, assuming linear income growth for simplicity, assign a mean income value for every age between 29 and 40 using this slope. For example, the estimated mean income value at age 30 is $(\overline{Income_{29}}) + 1 * (29\text{-}to\text{-}40\,slope)$.[16]

The process of estimating income at ages before age 29 and after age 40 is slightly more complicated. Since earnings growth flattens and starts to decline as workers age, we are not comfortable extrapolating the 29-to-40 slope beyond that age range. Using the 2011 Current Population Survey, we obtain three slopes between average family incomes at different ages: 22 to 29, 29 to 40, and 40 to 62. We then calculate two ratios: the ratio of the 22-to-29 slope to the 29-to-40 slope and the ratio of the 40-to-62 slope to the 29-40 slope.[17] We apply these ratios to the observed 29-to-40-slope in our SGM data to get estimated 22-to-29 and 40-62 slopes for our data. The two estimated slopes are used in the same way as the actual 29-to-40 slope to get income values for ages 22 to 28 and 41 to 62. For example, the estimated mean income value at age 41 is $(\overline{Income_{40}}) + 1 * (40\text{-}to\text{-}62\,slope)$.

---

[15] Note that we do policy simulations that include income-to-needs at age 29 and age 40 separately from the simulations that include income measured continuously in dollars. We consider income-to-needs solely in order to construct the success measures for TTA and adulthood. The basic simulation equations do not include income-to-needs, and the simulation equations to predict income-to-needs do not include income.

[16] We use mean incomes to compute the slope—as opposed to using individual incomes to compute individual-specific slopes—because some individual slopes are negative, which would complicate the estimation of stylized lifetime income effects. At the same time, our "spline" estimation prevents us from having to assume a linear growth rate, which would involve substantial under- and over-prediction of income at different points in the age profile.

[17] The ratio of 22-to-29 family income to 29-to-40 family income in the CPS is 1.70; the ratio of 40-to-62 income to 29-to-40 income is -0.19.

Each income (age 22, age 23, … , age 60) is discounted from birth using a real discount rate of 3%. So discounted age 40 income is $\frac{\overline{Income_{40}}}{1.03^{40}}$. Finally, lifetime family income is the sum of every discounted income:

$$discounted\ lifetime\ income = \sum_{i=22}^{62} \frac{\overline{Income_i}}{1.03^i} \qquad \text{Equation 5}$$

To estimate the *change* in lifetime income that results from an intervention or "what-if," this process is done with both pre- and post- income values. We subtract discounted lifetime income *pre* from discounted lifetime income *post* to get the mean change in lifetime income.

Challenges

There are a number of data and modeling challenges in building the model described above. Not only is there no perfect data set for this work, but modeling the life course is extraordinarily complicated. No model can ever fully capture all of the complexities of reality. The following section outlines the nature of these challenges and our methods for handling them.

Data Challenges

Data Availability and External Validity

Ideally, the SGM would be based on a single longitudinal dataset that follows individuals from birth to age 40 with no attrition or missing data. Unfortunately, no such dataset exists. All longitudinal datasets have item non-response and attrition from the survey. No American dataset follows a nationally representative group of children from birth through adulthood and includes reliable academic, cognitive, and behavioral measures at multiple ages. Faced with the necessity of linking multiple datasets, our primary goal was to use as few as possible in order to minimize the error that linking creates. That put a premium on finding two longitudinal datasets that together covered the entire period from birth to forty. Given these requirements, our choices for data narrowed quickly.[18]

The requirement that our data follow children over several decades presented an unavoidable dilemma: any real-world dataset following people over lengthy periods can accurately represent today's adults but not necessarily today's children. If a dataset includes contemporary adults who have been

[18] While the shortcomings of the CNLSY are unfortunate, it turns out that they are avoidable only at significant cost. While there are attractive alternatives to the CNLSY for early and middle childhood data, there is no satisfactory way to link middle childhood to adolescence without it. Therefore, fixing the problems noted above by resorting to another survey would come at the expense of having to add another "link" to the final birth-to-forty dataset. A table of 21 alternative datasets we considered is available on request. Several formal and informal advisors suggested we use the Panel Study of Income Dynamics (PSID) as the basis for our dataset. However, cognitive and behavioral outcomes in childhood are available only in the PSID's Child Development Supplement (CDS), which has three shortcomings for our purposes. First, the CDS has only been administered since 1997, which means that the oldest children with early childhood outcomes are only in their mid-20s. Second, there have been only three waves of the CDS spanning ten years. No child is observed in three consecutive stages. Finally, the sample sizes are too small. For instance, there are just 3,500 children in the 1997 wave, who are scattered across the ages of 0 to 12. Nor is the PSID likely to be better than the NLSY79 for imputing adult outcomes to CNLSY children. While we would get more recent data for 29-year-olds and for 40-year-olds using the PSID, there are fewer adolescent variables in common between the CNLSY and PSID than between the CNSLY and NLSY79, making imputations and simulations more problematic.

followed throughout their lives, then the data for earlier ages will be less informative about today's children. Today's children, for example, are much more diverse, much more likely to grow up with a single parent, and more likely to have working mothers than today's adults were as children.

On the other hand, a dataset of contemporary children has the problem that they will not be adults for some time. Assessing how children born in recent years will turn out at older ages requires extrapolating into an uncertain future, and the researcher must impute outcomes for each child for those older ages. The data used for imputations necessarily will come from earlier birth cohorts. In other words, this approach requires the assumption that today's children, when grown up, will resemble today's adults. But of course, much will change in the next forty years, potentially including educational attainment levels and the pay that people with different amounts of education will receive.[19]

Imputation

In both the NLSY79 and the CNLSY datasets, there is missing data due to non-response and attrition, and in the CNLSY data, there is missing data due to the censoring of children born before 1980 (who were too old at the start of the CNLSY in 1986 to have early childhood data) and born after 1990 (who were too young in 2010 to have adolescent, middle childhood, or even early childhood data, depending on their birth year).[20]

We impute values to missing data for all children ever observed in the CNLSY and all adults in the NLSY79, including non-responders and attriters as well as children in the CNLSY censored between birth and adolescence. We do so by first filling in values where we can by using a child's or adult's own non-missing data recorded for the same variable at some age close to the one with a missing response. For example, maternal education might be missing in the year of a child's birth but observed when the child was two.[21] About 90 percent of our CNLSY sample has at least one missing variable modified through this "proximity imputation" process, as does 72 percent of our NLSY79 sample.[22] Only 28 percent of our CNLSY sample has more than five variables with proximity imputations. On a variable-by-variable basis, between 0 and 40 percent of values are imputed in this way (see Table 5).

[19] Because the CNLSY children were born over a long period of time (one was born in 1971, and 3 were born in 2009, the most recent wave of data), they have experienced a diversity of experiences tied to different historical periods. This feature could be viewed as a problem in that any economic or societal changes affecting mobility could make the experiences of the children born in earlier years less relevant to the mobility of contemporary children. On the other hand, to the extent that we want to think of our data and model as applying to a sort of timeless set of children, in acknowledgment of our inability to fully predict what the future America will look like, using such a diverse group will help isolate the more general factors affecting mobility.

[20] The literature examining attrition in the CNLSY suggests that while it is non-random, the bias it introduces is not large. See Aughinbaugh (2004), London (2005), Cheadle, Amato, and King (2005), and Keng and Huffman (2007).

[21] In some cases we use an average of observed values at multiple ages or interpolate between ages. In other cases, we draw from values observed at the nearest-possible ages before successively looking for values at incrementally more-distant ages.

[22] Here we count a value as imputed if it was drawn from an age other than 0 for birth, 5 or 6 in early childhood, 10 or 11 in middle childhood, 29 or 30 in transition to adulthood, or 40 or 41 in adulthood. Because the CNLSY is biennial and children may be interviewed before or after their birthday, depending on the time of year, some children end up being (for example) 7 years old rather than 5 or 6 when they are interviewed, which would count as an "imputation" for the above purposes.

After this initial imputation, we then impute remaining values using linear or nonlinear models applied to non-missing data to predict values for missing data.[23] In the CNLSY, we start with our birth circumstances variables and order them from the one with the least missing data to the one with the most missing data. One by one, we predict each variable from more-complete ones.[24] By ordering variables according to how many missing values they have, we minimize the extent to which imputations are based on other imputed values. Once all of the birth circumstances variables have been imputed in this way, we move to the early childhood variables, beginning again with the one that has the least missing data and predicting it from the birth circumstances variables. We continue in this way until we have a completely filled-in birth-to-19 dataset. We then iteratively impute missing values in the same way in the NLSY79 to build a 19-to-40 dataset.[25]

The extent of our model-based imputation for each variable is given in Table 5. In the CNLSY, just 21 percent or fewer observations have model-based imputations for each at-birth variable, except that 51 percent of PPVT (vocabulary) scores are imputed in this way. The prevalence of model-based imputation rises to 15 to 25 percent for early and middle childhood variables, and then 12 to 78 percent for adolescent variables. This steady increase reflects the issue of censored children born to older mothers, whose later outcomes have not yet been observed. Model-based imputation is much less common in the NLSY79, with the exceptions of the grade point average variable and many of the behavioral variables used to link the two datasets and the adult income variable (our ultimate outcome). The conviction variable and many of the adolescent behavioral variables also have high levels of missingness because those questions were only asked in a single year. Table 5 also includes the R^2 or pseudo- R^2 values for the models used to impute missing values for each variable. The statistic provides

[23] These imputation models include ordinary least squares, logit, ordered logit, and multinomial logit models. Prior to model-based imputation, we drop a very small number of observations in each dataset missing data on race (3 in the CNLSY, 16 in the NLSY79).

[24] For example, we start with maternal age at child's birth (one missing value), which is predicted from race and gender as well as from maternal age at *first* child's birth and birth order, which have no missing values. The predictions from the model are used to replace all missing maternal-age-at-birth values (in this case only one). Next is maternal education (59 missing values), predicted from race, gender, maternal age at first birth, maternal age at child's birth, and birth order.

[25] We also incorporated variability into our imputed values. For continuous measures, we did so by randomly drawing from the distribution of residuals in our imputation models and assigning one to each case to which we imputed values. When imputing values to categorical variables, we compared predicted probabilities against random draws from uniform distributions in order to assign cases to one category or another. We tested this "stage-by-stage" imputation procedure against a second approach that links together the sequence of imputation models and then iteratively improves the imputations *across* the stages by updating them based on what the values would "most likely" be given the patterns in the observed data. In this alternate approach, even later-stage variables are used to impute values to variables with missing data. Technically, this check uses a "multiple imputation by chained equations" (MICE) algorithm, which is a Bayesian simulation algorithm that uses observed data to generate posterior distributions of missing data, the values of which are then used to replace missing data (van Buuren and Oudshoorn 1999). Note that we do not rely more generally on multiple imputation, which can improve the variance estimates produced by statistical analyses in the presence of missing data. The main is reason is that to date, incorporating variability into our results has been a secondary concern. The systematic error in our data is likely to swamp the classical error, so conventional methods to assess variability that focus on sampling error are less appropriate than they would be for simpler datasets. Comparing results using both approaches reassured us that our home-grown strategy produced valid estimates.

an indication of how valid our imputations are likely to be, where a value of 1.00 would indicate perfect prediction and a value of 0.00 a prediction no better than randomly assigning values.[26]

Our decisions around missing values mean that our data include a substantial amount of imputation. In the CNLSY data, for instance, all of the children have at least one outcome imputed using modeling. Sixty-five percent have at least five model-imputed values, 49 percent have at least nine, and 30 percent have fifteen or more. In the NLSY79, 90 percent of adults have at least one model-based imputation, and half have more than two.

In the CNLSY, over 40 percent of our sample consists of children with censored data—born before 1980 or after 1990 and therefore unobserved either in early life stages or in later ones—and so entire life stages are imputed for them.[27] (In comparison, about one-fifth of our sample consists of non-censored children—old enough to be observed from birth through age 19— who attrited.) In earlier stages of our research, we conducted analyses that excluded these censored children and also discarded children and adults with more than five model-based imputations. We were, however, dissatisfied with not having a broadly representative sample of children for descriptive analyses or simulations.[28]

Figure 2 illustrates the extent to which the sample of non-censored children disproportionately consists of children born to relatively young mothers. Comparisons with the full CNLSY indicated that the sample of non-censored children was disproportionately comprised of racial and ethnic minorities and was distinctly disadvantaged compared to all CNLSY children on measures like maternal education, family structure, and income at birth. There were also large differences in life-stage-specific success rates between the non-censored children and the full set of children when we constructed our success indicators.

Projecting Adult Outcomes

The filled-in CNLSY allows us to follow children from birth to age nineteen. The filled-in NLSY79 lets us track a different group from age nineteen to age forty. The remaining challenge was to determine how to use the NLSY79 data to impute, or project, post-adolescent outcomes for the CNLSY children so that we can "follow" them from birth to forty. We did so using microsimulation to predict CNLSY outcomes based on the relationships between variables in the NLSY79.

As discussed above, we estimate models predicting transition-to-adulthood outcomes using the NLSY79. As regressors in these models, we use only at-birth and adolescent variables that are available

[26] The family income variables have such high R^2 values because they are predicted from, among other variables, income to needs. The low values for the dichotomous variables reflect the well-known problem that linear probability models under-estimate R^2 values (see Greene, 1981, 1983). A second measure of the quality of our predictions is given in the table—the correlation (across non-imputed observations) of observed values and predicted values.

[27] Children born before 1980 (who are not observed in early childhood) are 11 percent of our final sample, while children born after 1990 (who are not observed in adolescence) are 31 percent of the sample.

[28] The alternative to imputing values explicitly, for purposes of creating a sample that is entirely non-censored and nationally representative, is to re-weight the existing non-censored data. This is simply an implicit form of imputation, however, and requires the assumption that within the strata for which weights are recalibrated, censored children will have the same outcomes as non-censored children (despite being born to relatively young or old mothers).

both in the NLSY79 and the CNLSY.[29] We then apply the coefficients estimated from the transition-to-adulthood models to the CNLSY variables that are common between both datasets to simulate transition-to-adulthood outcomes for the CNLSY children. For each child and each outcome, we add "unexplained" variation by giving them error terms based on each model's standard error of regression.

Finally, we estimate a model predicting income in adulthood, again using only the NLSY79 data. We predict it from the transition-to-adulthood outcomes in the NLSY79 and from the prior variables that are in both the NLSY79 and CNLSY. Then we apply the results to the simulated transition-to-adulthood outcomes we created in the previous step (also using the common at-birth and adolescent variables) to get simulated income at age 40. The basic assumption in this approach to imputing values is that the relationships observed between variables in the NLSY79 are similar to the relationships we would observe between the same variables if they were available in the CNLSY.

Benchmarking

Given the inherent challenges in linking datasets and relying heavily on imputation, we were particularly concerned with verifying that our final dataset plausibly represents a contemporary and representative group of children and the experiences they will have as they become adults. Fortunately, the evidence we have assembled using outside data sources has reassured us that on a number of dimensions, our data hit appropriate targets quite well. The notable exception is that the associations between our early and middle childhood test scores, and between those scores and several demographic variables, appear to be lower than they should.

Demographics and Early Childhood Income

The children in the CNLSY were born to parents who were in the country in 1979, and of the cross-sectional NLSY79 sample, only about 4% of youth in the parent generation were born outside of the U.S. This means we can think of our sample as representing native-born children of native-born adults. The Urban Institute's Children of Immigrants Data Tool uses American Community Survey (ACS) data and allows users to look at the racial composition of native-born children of native-born adults. The data from 2005 to 2006—the earliest available—indicate that among children under eighteen (born 1988 to 2006), 71% were white, 18% black, 10% Hispanic, and 2% "other". In our dataset, where most children were born in the 1980s and 1990s, the corresponding figures are 71%, 14%, 11%, and 4%. The small discrepancy is likely just a result of the different definitions of the categories; our "other" category includes people of more than one race, while the Children of Immigrants Data Tool reports only the Asian and Native American population as separate from white, black, and Hispanic.[30]

We also compared the means for several other variables in our data to those in the Department of Education's Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K).[31] The ECLS-K children

[29] Specifically, we predict transition-to-adulthood outcomes from the adolescent variables shown in Table 2 and from race, gender, maternal education, and maternal age.

[30] The Urban Institute Children of Immigrants Data Tool Technical Appendix states that "Non-Hispanic blacks are all those who reported they were black or African American, regardless of additional racial/ethnic groups [aside from Hispanic] reported." Presumably some individuals who get counted as non-Hispanic black by Urban would fall into our "other race" category.

[31] We are indebted to Katherine Magnuson of the University of Wisconsin-Madison for the ECLS-K tabulations. Note that the ECLS-K does confirm that including children of immigrants and immigrant children—and looking at a

were born primarily in 1992 and 1993, and they represent kindergarteners in 1998. Comparing our sample means to the ECLS-K means—a sample that includes children of immigrants and immigrant children—mothers in our sample were, on average half a year younger when their children were born and half a year younger when they had their first child. While 8 percent of children in our sample were born low birth weight, 7 percent in the ECLS-K were. In early childhood, average income to needs among the children in our data was 2.96, while it was 2.86 in the ECLS-K. In our data, 19% of children were poor in early childhood, the same proportion as in the ECLS-K. And the share of children living with two married parents in early childhood was 73% in both datasets.

Finally, cross-tabulations of race and poverty status match quite well between our data and the ECLS-K (see Figure 3). Whites and blacks in our data are slightly less advantaged and Hispanics more advantaged (as we would expect since SGM children are all native born).

Childhood Academic and Behavioral Skills [32]

The ECLS-K measures many early and middle childhood outcomes which are similar to the SGM childhood outcomes. It administers math and reading achievement tests and asks parents and teachers questions about their children's behavior and social skills. We cannot generally compare the means of these outcomes between the two datasets because the variables are measured on different scales. Therefore, we focus on comparing the strength of relationships between the various childhood outcomes, as well as achievement gaps between socioeconomic groups in each dataset.

Table 6 shows correlations between math and reading achievement within early and middle childhood and across the two life stages. Associations between achievement test scores are uniformly and substantively lower in our dataset than in the ECLS-K. The tests used in the ECLS-K were better designed than those in the CNLSY; they have more items and rely on sophisticated item response theory (IRT) methods. The test reliabilities are much higher than in the CNLSY. Furthermore, the PIAT tests in the CNLSY were designed for and normed on children in school in the late 1960s. Because of the well-known "Flynn Effect," named after psychometrician James Flynn, which describes increases in test scores over successive cohorts of children given the same test, the CNLSY children do better on the PIATs than the tests assume should be the case.

Test reliability differences are also likely to blame for the fact that test score gaps in the SGM dataset are quite a bit smaller than in the ECLS-K (see Figure 4). Once again, the exclusion of immigrant children and children of immigrants surely contributes to the under-estimation of gaps between Hispanics and whites, but the other gaps are also underestimated (as are gender gaps, which are much smaller in both datasets, and gaps according to family structure and maternal age at birth).

In contrast to the low correlations between test scores in the SGM dataset, the associations between behavioral outcomes are fairly close to those observed in the ECLS-K (Table 7). This similarity is surprising, because the behavioral measures are, if anything, less consistent across the two datasets

---

more recent birth cohort—does affect the race/ethnicity distribution of children. In the ECLS-K, just 57% of children are white, 16% are black, 19% are Hispanic, and 8% are "other". But Magnuson found that the means for the other variables examined did not change a lot when the sample was restricted so that it excluded children of immigrant mothers and the youngest and oldest mothers.

[32] We gratefully acknowledge Magnuson again, who conducted the ECLS-K analyses in this section.

than the achievement tests. For example, in our data, the measures are reported by parents, while in the ECLS-K they are reported by teachers.[33]

Educational Attainment

We measure educational attainment by the percentage of respondents who report having obtained a high school diploma by age 19 and by the percentage of respondents who report having obtained a bachelor's degree by age 29. The majority of our sample was born between 1980 and 1990. Those children turned 19 between 1999 and 2009, and turn 29 between 2009 and 2019. To look at high school graduation rates for a roughly comparable cohort, we looked at rates for 2005. To look at college graduation rates for a roughly comparable cohort, we looked at 2011 rates.

The Census Bureau reports high school graduation rates from the Current Population Survey (CPS) by counting GED holders as having graduated. Therefore, for benchmarking purposes we report graduation rates in the SGM data using the same definition. Figure 5 shows the percentage of the population age 20 to 24 with a high school degree, according to the CPS, alongside the estimates from the SGM data for 19-year-olds. Our graduation rates are a bit higher than the CPS rates, even though the CPS figures are for men and women who have had one to five years longer to get their degree than the people in our data. The big exception is that we show much higher graduation rates for Hispanics, which is almost surely driven by our sample's omission of immigrant children and children of immigrants.

We also turned to the CPS to benchmark college completion rates. In many ways, this is a more important check than the ones mentioned to this point. That is because college graduation rates in the SGM dataset are based on simulated values, as discussed above. Benchmarking here is a bit more speculative as well, because the children in our sample are not yet old enough as a group to have passed through the college-going years—we must project what their college graduation rates will look like. The SGM measures whether a person has a bachelor's degree or higher by age 29; we compare our estimates against published 2011 CPS figures for percentage of the population age 30 to 34 with at least a bachelor's degree. Our predicted rates of college graduation tend to be somewhat lower than those observed in the CPS. The extent of our under-prediction is probably understated to the extent that future college graduation rates will be higher than today's. Again, we over-predict the Hispanic rate.

Income

Table 8 compares our family income and income-to-needs estimates to benchmarks from several datasets. We compare income at birth to estimates from CPS microdata for children under one year old in 1988 (family income measured in 1987, the median birth year in our SGM sample).[34] As Table 8 shows we do very well against the CPS benchmarks for both income and income-to-needs.

Next, we compare our income-at-40 estimates against two benchmarks: the CPS (focusing on 2002 family income among families with a head who was between 38 and 42 in 2003, by which time most of the NLSY79 sample members had turned 40) and the NLSY79. This check is especially important

[33] It is not straightforward to standardize the behavioral problems measures in a way that allows for gaps between groups to be compared meaningfully.

[34] We manipulate CPS data in several ways to make it comparable to SGM data. We transform everything to 2010 dollars; we topcode income at $156,000 and income-to-needs ratio at 13; and we define "family income" to encompass single people.

because in the SGM dataset, age-40 incomes are simulated. Once again, our SGM means track estimates from these two sources very well.

Finally, we use the Panel Study of Income Dynamics (PSID) to look at the adult family incomes of people who were born into poverty. This last check is important because it allows us to assess how good our simulated adulthood data is for a subgroup defined by the (non-simulated) at-birth data. In the PSID, the same people have been followed from birth to forty. The comparison is not ideal because adults in the PSID sample were around 40 years old in 2006 or 2008 (the years examined) and were born in the late 1960s and early 1970s, so their experience may differ from the experience of more recent birth cohorts.[35] Our income-to-needs estimate for adults who were poor at birth is reasonably close to the PSID benchmark—3.0 vs. 3.6 in the PSID—but our income estimate is lower by $13,000. It is difficult to interpret how problematic this discrepancy is given the cohort differences and possible discrepancies between the PSID and SGM analyses. Reassuringly, our median income and income-to-needs estimates matched the PSID ones more closely.[36]

Modeling Challenges

A model is always a simplification of real-world complexities, but to be useful, it must do a reasonably good job of characterizing reality. In this section we address four threats to the internal validity of the SGM: missing mediators, incompletely-modeled interventions, measurement error, and omitted variable bias.

Missing Mediators

A potential threat to the validity of our model is the possibility that there are mediating (intervening) variables missing from the model. To understand why this might be a problem, it is useful to distinguish between direct effects and indirect effects. Through adolescence, any X in the SGM can affect any Y either directly—as estimated by the coefficient on the variable in a regression equation—or indirectly through mediating variables. For instance, maternal education can affect middle childhood test scores directly because it is included in our equations predicting test scores. It can also affect them indirectly through the four early childhood outcomes. Maternal education directly affects those early childhood outcomes, which then affect middle childhood test scores. The sum of the direct effect and all indirect effects equals the total effect of an X on a Y. That is the relevant effect in microsimulation—the total amount Y changes in response to X.

If we had a single birth-to-forty dataset, we would not have to worry about missing mediators. Adding mediating variables simply shifts the way total effects are allocated between direct and indirect effects. For example, adding a fifth early childhood outcome would change the direct effect of maternal education on middle childhood math scores, but it would not change the total effect. By the same logic, we do not have to worry that we are missing some early childhood outcome that mediates the effect of maternal education. An intervention that increases maternal education will fully propagate forward to raise middle childhood math scores regardless of the specific pathways it takes to get there.

[35] We are indebted to Kathleen Ziol-Guest of Cornell University and Greg Duncan of University of California-Irvine for the PSID analyses.

[36] One reason PSID means are higher may be due to top-coding, as the PSID analyses were conducted independently of ours.

However, the seam in our data at adolescence—where the CNLSY ends and where we have to simulate subsequent outcomes via microsimulation and the NLSY79—disrupts this convenient property of the model. Because our regressions estimating TTA outcomes include only adolescent and at-birth variables that are available in both the CNLSY and the NLSY79, interventions in early and middle childhood (and interventions on at-birth variables not available in the NLSY79) cannot propagate fully forward to TTA. Consider an intervention on early childhood reading scores. The effect will fully propagate through the four middle childhood outcomes and through the 19 adolescent outcomes. But the direct effect of the improved early childhood reading scores on TTA outcomes cannot be modeled, nor can the indirect effects that go through middle childhood outcomes and then bypass adolescent outcomes (because there are no direct effects from middle childhood outcomes on TTA outcomes). The problem is even worse for estimating effects on adulthood income because of missing direct effects from early childhood and middle childhood to both TTA outcomes and adulthood income. Any effects from early or middle childhood (or from at-birth interventions on variables not in the NLSY79) propagate only through adolescent outcomes in our model.

When we added three adolescent outcomes to the model, it increased the estimated effects of early interventions on TTA outcomes and adulthood income by a sizable amount—tripling, for instance, the effect of an early childhood intervention on adult income. On the other hand, adding 12 additional variables did not meaningfully change our estimated effects significantly, suggesting that we have largely addressed this problem.

Incompletely Modeled Interventions

Missing mediators across our seam constitute an "omitted pathways" problem. A second such problem exists when we cannot model all of the pathways whereby an intervention affects later outcomes. There are two situations in which this might be the case. First, it may be that there is no good evidence on how some intervention affects one of the variables in our model. We might wish to model how some program to improve parenting affects the later outcomes of children, but the program evaluation from which we draw initial effect sizes may not have assessed how the intervention affected antisocial behavior. In that case, if we assume the effect of the parenting program on antisocial behavior is zero, we will have an omitted pathway from the intervention to later outcomes, and the long-term effect of the program is likely to be understated.

The second situation in which we might incompletely model an intervention's effects is if some program evaluation determines an effect on some outcome that we do not have in our model. Perhaps the parenting program boosted child self-esteem, which we might expect would eventually increase the likelihood she will attend college. Our model lacks a measure of child self-esteem, and so this will be another omitted pathway whereby the intervention has an effect on long-term outcomes. Our estimated long-term effect will again be understated.

To assess the extent to which such omitted pathways could be an issue for the SGM, we reviewed a number of long-term studies of early interventions that measure total effects. We also commissioned research to compare the estimates our model generated for specific early interventions for which long-term effects have been evaluated to the actual results found in the evaluations. That is, we have tried to determine whether the explicit pathways we model generate long-term effect

estimates comparable to studies that have measured total effects without regard to the pathways. We turn to this benchmarking in the section below, "Validating the Model."

Measurement Error

Measurement error is the difference between the true value of some variable for some person and the measured value. The variables in our data suffer from two kinds of measurement error—random and systematic. In turn, both random and systematic error might be due to error in the variables we observe or to error introduced when we impute values to missing data.

Random measurement error occurs when the error in the variable is independent of the distribution of other variables. If a test-taker scores slightly higher or lower than he would have had he taken the test the day before, with these daily fluctuations distributed around his "true" performance in some pattern-less way, that constitutes random measurement error. If our imputations assign values that are sometimes too high, sometimes too low, but not in any systematic way, that too is random measurement error. Random measurement error acts as a downward pull on our estimated effects, causing them to be understated. Systematic error occurs when errors are patterned in some way that is related to other variables. Systematic error can bias effects upward or downward.[37]

It is difficult to assess the extent to which our imputations might exacerbate any measurement error problems we might have in the absence of missing values. For the most part, as is generally true in social scientific analyses, our predictive equations have fairly low explanatory power (see Table 5), and the relationships between heavily-imputed variables in our model and subsequent outcomes may be attenuated. We believe that the most serious issue is the systematic error that is introduced by simulating TTA outcomes and adulthood income. However, since we cannot know what the "true" values of these outcomes will be for the CNLSY youth in the future, it is not possible to assess the extent of error beyond the benchmarking described above.[38]

Omitted Variable Bias

Any exogenous variables that are missing from our model could lead to the well-known problem of omitted variable bias (OVB). To illustrate OVB, imagine that we have left out some CAB variable that affects both early childhood and middle childhood math scores directly. The effect of early childhood scores on middle childhood ones that we estimate will be biased—probably upward (if the omitted CAB

[37] A reading test that is biased against children of a certain background—that assesses their true reading ability less well than it does for other children—would produce systematic error. Imputations that are uniformly too high or too low would constitute another form of systematic error, which might be the case if people with missing values are systematically different than those with observed values in ways not captured by the imputation regressions. Another example where that might be the case would be if the NLSY79-based coefficients we use to simulate TTA outcomes and adulthood income for CNLSY youth do not reflect the real-world associations those youth will actually see (e.g., the return to a high school or college degree might have increased). Finally, we might underestimate the levels of income in TTA and adulthood through our NLSY79-based simulation because economic growth could push average income levels up relative to those seen by the NLSY79 cohorts—the intercept from the NLSY79 could be too low.

[38] We have confirmed that the primary alternative to simulating these outcomes—statistical matching of NLSY79 adults to CNLSY youth—produces demonstrably implausible results. Imputations aside, we did attempt to incorporate errors-in-variables corrections in our regressions of outcomes on several variables for which published reliabilities are available. Unfortunately, when making such corrections for multiple variables with relatively low reliabilities, estimation problems arise, preventing us from implementing the correction.

variable is related to both scores in the same way). When we then simulate some intervention that raises early childhood math scores, because of the OVB, our estimated effect of the intervention on middle childhood scores—operating through early childhood scores—will be too big. A crucial point in understanding whether OVB is a problem for the SGM, however, is to distinguish between an individual coefficient or effect being biased and the *total* effect of the intervention being biased. Only the latter is a problem, and only some kinds of OVB will bias our total effects.

Understanding omitted variable bias is complicated in the context of a recursive system of equations such as that constituting the SGM. There are two clear instances when OVB potentially biases our total effects. First, OVB is potentially problematic for us when omitted variables are temporally prior to the intervention stage and they affect subsequent variables that are targets of the intervention. In the math score example, the omitted CAB variable will bias the coefficient on the early childhood score in the equation predicting the middle childhood score. As a consequence, that biased effect will propagate forward to estimates for subsequent-stage effects; the too-high middle childhood math scores will lead to too-high GPAs in adolescence, for example. In addition, the direct effect of early childhood math scores on adolescent variables will also be biased upwards.

On the other hand, for purposes of estimating total effects on early childhood, middle childhood, or adolescent outcomes, we do not believe that omitting variables that come during or after the intervention stage biases our estimates. Imagine that we simulate an intervention that increases parental income at birth but that we omit some variable in early childhood that affects middle childhood math scores and GPA in adolescence directly. Clearly, this is a classic instance of OVB where the coefficient on middle childhood math scores in the equation predicting GPA will be biased. What is less obvious is that by conditioning on middle childhood math scores, we create an association (or an additional source of association) between the omitted variable and the other early childhood variables in the GPA equation, which all affect high school GPA directly in our model. That will bias the coefficients on the included early childhood variables. But it turns out that the total effect of the at-birth intervention on high school GPA will be unbiased because the biases in these coefficients are offsetting.[39] (All of this assumes that we have correctly modeled all of the direct effects of the at-birth intervention, as discussed above, and that we have not omitted pre-birth variables that lead to OVB.)

The key here is that we include direct effects from early childhood variables when predicting adolescent variables, which allows biases in individual coefficients caused by omitting an early childhood variable to offset one another. However, when predicting post-adolescent outcomes, the necessity of using the NLSY79 prevents us from modeling the direct effects of EC and MC variables (and some CAB variables) on TTA and adulthood outcomes. That means that omitting variables that come after the

[39] In the evolving literature on "directed acyclical graphs" and causal diagrams, this is known as "conditioning on a collider." (Pearl, 2000) To illustrate the idea, imagine that one can only get a good job by being male or by being a college graduate, and that gender, education, and one's job all affect income directly. Also imagine that there is no correlation between being male and being a college graduate. Despite this absence of association, conditional on having a good job there will be a negative association between being male and being a college graduate. If we predict income from gender and whether or not someone has a good job (omitting whether or not someone is a college graduate), we clearly bias the coefficient on having a good job. But we also bias the coefficient on gender, because we have controlled for having a good job, thereby inducing an association between gender and income through the induced negative association of gender and (the omitted) college graduation variable.

intervention stage *can* bias total effects of pre-adolescent interventions on post-adolescent outcomes—the second case where OVB is clearly a potential problem for the SGM.

To address this second case, we have focused on including as many adolescent variables as possible in our model, which will soak up some of the OVB from having left out pre-adolescent (but post-intervention) variables. Adding variables to adolescence, of course, also helps soak up OVB when we simulate an intervention in adolescence or TTA. We confirmed that adding fifteen adolescent variables to our model made a noticeable difference on our estimates when we simulated an intervention in adolescence, cutting our estimated effects on adult income in half.

To address omitted variables prior to the intervention stage (as well as address missing mediators), we added several CAB and pre-EC variables to our model, including parenting measures, a measure of maternal cognitive skill, and a vocabulary test score from age 3 or 4. It appears to have had a relatively modest impact on our long-term estimated effects.

We have also explored the likely importance of OVB by estimating a version of the SGM in which we controlled for family fixed effects in early childhood and simulated early childhood interventions. That is, in the equations predicting middle childhood outcomes, we included indicator variables for the CNLSY family in which a child lived. The fixed effects control for all shared influences—genetic and environmental—between children living in the same family. Our estimated effects using this approach were only a bit smaller than in our standard model.[40]

Finally, we have explored the extent to which our estimate of a key relationship in our model compares with estimates from the literature that attempt to address omitted variable bias in a rigorous way. The relationship between education and adult economic outcomes is at the core of our model. Estimating the effect of schooling on future income, however, is a classic instance where OVB might bias the results; people advantaged in terms of some unobservable quality such as ability or motivation may tend to obtain more schooling and also tend to earn higher incomes, without their educational attainment actually being important.

A number of studies examining the relationship between earnings and educational attainment have used quasi-experimental identification strategies (see Ashenfelter and Rouse, 1999 and Card, 2001 for reviews). They find that an additional year of schooling is worth earnings boosts of anywhere between 3.7 percent and 13.2 percent.[41] Card (2001) finds that one of these identification strategies, using "instrumental variables," tends to yield bigger estimates of the effect than conventional estimates, which is the opposite of what one would expect if omitted variable bias were a problem. Because research has determined that random measurement error in schooling reports biases estimated effects of schooling downward by 10 to 40 percent, Card (2001) and Ashenfelter and Rouse (1999) have speculated that in practice, for this particular relationship, OVB and measurement error may effectively cancel out. Our own checks using the NLSY79, comparing earnings and educational attainment in the

[40] This exercise likely understates the importance of OVB to our results in that it does not control for omitted variables that differ within families (such as personality traits) and in that it does not include family fixed effects in predicting outcomes in adolescence and later. Since early childhood interventions can affect things that siblings share, including fixed effects in subsequent stages is arguably over-controlling.

[41] This includes a range of 3.7 to 6.3 percent from sibling fixed effects models, an average of roughly 8 percent for studies including identical and fraternal twins, and a range of 6.0 to 13.2 percent for instrumental variables studies.

same year and controlling for a number of variables available in adolescence and at birth, returned an estimated effect of an additional year of schooling on earnings of 7 percent.

Validating the Model

Our attempts to validate the long-term effects estimated by the SGM have yielded mixed results. On the one hand, recent research by Raj Chetty with various colleagues to estimate effects of early educational interventions on long-term outcomes has encouraged us (Chetty et al., 2011; Chetty, Friedman, and Rockoff, 2011). In the first paper, Chetty and his colleagues link tax data from the IRS to young adults who participated in the Project STAR class-size experiment as children. Children in the Tennessee experiment were randomized into regular-sized classes the first year they entered elementary school (kindergarten or first grade unless they transferred in) or classes that were smaller by about one-third. Teachers within the schools were also randomized to a normal-sized or small classroom. Once randomized, children generally stayed in either a normal or small classroom through third grade. The experiment took place during the second half of the 1980s, and Chetty et al. link participants to their tax records during their mid- to late-twenties.

The researchers find that being initially assigned to a small class through third grade increased the probability of attending college by 1.6 percentage points and increased earnings by less than $500 (if at all). There were no significant effects on homeownership or marriage. They also find that having a kindergarten teacher with over ten years of experience raised earnings by $1,100, or 7 percent of the mean, however there was no effect for children entering the experiment after kindergarten. Having a teacher with a graduate degree did not affect earnings, and the effects of classmate demographic composition were also generally small and not statistically significant. Finally, a one-standard deviation in class quality, as measured by classmates' test scores, raised earnings by $455 to $1,520 (about 3 to 10 percent of the mean).

Chetty, Friedman, and Rockoff (2011) consider teacher value-added—the effect of a teacher on student test scores—and whether it affects students' adult outcomes. They use administrative data from a large urban school district and tax data from the IRS. The authors find, for instance, that raising the value-added of a student's teacher in a single grade by one standard deviation increased earnings at age 28 by 1 percent. The results from these papers are in some sense discouraging from the perspective of public policy, in that relatively large interventions are shown to have relatively modest long-term effects. But they accord reasonably well with the magnitude of estimated effects typically produced by our model.

For example, working with the RAND Corporation's Lynn Karoly, we simulated the long-term effects of three widely-admired early childhood programs: the Carolina Abcedarian program, the High/Scope Perry Preschool Project, and the Chicago Child-Parent Centers program. We simulated outcomes among poor children born to young mothers and who had low preschool vocabulary scores, our best attempt to define a target population comparable to the children served by these programs. We increased early childhood math and reading scores in accordance with the short-term effects actually produced by the programs.[42] The model estimated that each program would raise college

[42] The Abcedarian and Perry Preschool studies were randomized experiments, while the Chicago CPC evaluation was based on a quasi-experimental design.

graduation rates among this group by two percentage points and increase income at age 40 by $2,000 to $3,000 (or about 5 to 7 percent).

The increase in college graduation was comparable to the actual effects produced by Perry Preschool and Chicago CPC, but age forty incomes are unavailable for the three studies. However, the SGM significantly under-predicted effects on middle childhood and adolescent outcomes compared with the actual effects the studies produced. For instance, the model predicted that high school graduation would rise by 2 to 3.5 percentage points but the studies showed actual gains of 10 to 18 points. It performed similarly poorly predicting effects on teen births and conviction rates and moderately under-predicted middle childhood math and reading scores.[43]

In defense of our model, it is parameterized mostly on a sample of children born in the 1980s and 1990s. The Perry Preschool Project was implemented during the 1960s and Abcedarian in the 1970s, and the Chicago CPC evaluation was conducted in the 1980s. The children participating in these programs were probably more disadvantaged than the target population we specified. Nevertheless, we would prefer to have come closer to the real-world results than we did.

Conclusion

The benefits of a model that can successfully explicate the processes underlying social mobility are readily apparent. Such a model would shed light on the basic patterns that typify the life courses of American children, providing policymakers with important information and suggesting avenues for further research. Perhaps more importantly, it would give policymakers information about the likely success of different approaches to promoting social mobility.

As should be apparent, developing a valid model of social mobility is no easy task—if it were, the need for such a model would have been filled by now. It is our hope that the Social Genome Model can inform policy debates and help to allocate scarce resources toward the widely-embraced goal of greater upward mobility. The alternatives are decision-making without information about long-term effects—or waiting a generation to observe long-term effects in actual evaluations of policy interventions.

[43] We were only able to model effects of the early childhood programs on achievement test scores. Heckman, Pinto, and Savelyev (2012) find that the strong effects of Perry Preschool on long-term outcomes was mediated primarily by noncognitive skills such as reduced externalizing behavior problems and greater academic motivation. When we specified arbitrary effects of the program on our two measures of externalizing behaviors, however, our estimates were unaffected.

References

Ashenfelter, Orley, & Cecelia Rouse. "Schooling, Intelligence, and Income in America: Cracks in the Bell Curve" (1999). NBER Working Paper no. 6902.

Aughinbaugh, Alison. "The Impact of Attrition on the Children of the NLSY79." *Journal of Human Resources* 39, no. 2 (2004): 536-563.

Card, David. "Estimating the Return to Schoolings: Progress on Some Persistent Econometric Problems." *Econometrica* 69 (5) (2001): 1127-1160.

Cheadle, Jacob E., Paul R. Amato, and Valarie King. "Patterns of Nonresident Father Contact." *Demography* 47, no. 1 (2010): 205-225.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR." *Quarterly Journal of Economics* 126(4) (2011): 1593-1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." (2011). NBER Working Paper No. 17699.

Duckworth, Angela L., and Martin EP Seligman. "Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents." *Psychological Science* 16, no. 12 (2005): 939-944.

Greene, William H. "Estimation of Limited Dependent Variable Models by Ordinary Least Squares and Method of Moments." *Journal of Econometrics* 21(2) (1983): 195-212.

Greene, William H. "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model." *Econometrica* 49(2) (1981): 505-515.

Heckman, James J., Rodrigo Pinto, and Peter A. Savelyev. "Understanding the Mechanisms Through Which an Influential Early Childhood Intervention Program Boosted Adult Outcomes." (2012). NBER Working Paper No. 18581.

Heckman, James J., and Yona Rubinstein. "The Importance of Noncognitive Skills: Lessons from the GED Testing Program." *The American Economic Review* 91, no. 2 (2001): 145-149.

Heckman, James J., Jora Stixrud, and Sergio Urzua. "The Effects Of Cognitive and Noncognitive Abilities On Labor Market Outcomes and Social Behavior," *Journal of Labor Economics* 24, no. (2006): 411-482.

Keng, Shao-Hsun and Wallace E. Huffman. "Binge Drinking and Labor Market Success: A Longitudinal Study on Young People." *Journal of Population Economics* 20, no. 1 (2007): 35-54.

London, Rebecca. "Welfare Recipients' College Attendance and Consequences for Time-Limited Aid." *Social Science Quarterly* 86, no. s1 (2005): 1104-1122.

Tyler, John H. "Economic Benefits of the GED: Lessons from Recent Research." *Review of Educational Research* 73, no. 3 (2003): 369-405.

Pearl, Judea. *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2000).

Shonkoff, Jack P., and Deborah A. Phillips. *From Neurons to Neighborhoods*. Washington: National Academy Press, 2000.

Van Buuren, Stef and Karin Outshoorn. *Flexible multivariate imputation by MICE*. Leiden, The Netherlands: TNO prevention and Health, 1999. Available at http://www.stefvanbuuren.nl/publications/Flexible%20multivariate%20-%20TNO99054%201999.pdf. Accessed September 14, 2012.

Tables and Figures

Table 1: Life Stages and Corresponding Outcomes

Stage	Variable	
Circumstances at Birth	Gender	A dichotomous variable indicating the sex of the individual. Males are the omitted category.
	Race	Dichotomous variables indicating whether the child is black, Hispanic, or other. The omitted category consists of white children.
	Maternal Educational Attainment	Dichotomous variables are included to indicate whether the individual's mother graduated from high school, attended some college, or obtained a Bachelor's degree or more advanced degree. The omitted category is mothers who did not finish high school.
	Maternal Age at the Time of the Child's Birth	A continuous variable measuring the age of the mother (in years) at the time of the child's birth.
	Maternal Age at First Birth	A continuous variable measuring the age of the mother (in years) at the time of her first child's birth.
	Marital Status of the Child's Parents at the Time of Birth	A dichotomous variable indicating whether the child's mother was married when he/she was born. The omitted category includes those children whose mothers were not married, even if cohabitating, at the time of their birth.
	Family Income at Birth	This continuous variable is the log-transformed measure of the family's income as a percent of the federal poverty line in the year that the child was born.
	Low Birth Weight	A dichotomous variable indicating whether a child weighed 5.5 pounds or less when they were born. The omitted category consists of children who weighed more than 5.5 pounds at the time of their birth.
	Mother's AFQT Score	The age-normed percentile score of the child's mother on the Armed Forces Qualifying Test, a general achievement test taken when the mothers were between 16 and 23.
	Parenting: Cognitive Stimulation	Standardized score on the HOME Inventory Cognitive Stimulation scale, measured when the child is 0-2.
	Parenting: Emotional Support	Standardized score on the HOME Inventory Emotional Support scale, measured when the child is 0-2.
	Early Verbal Ability	The age-standardized score of the child on the Peabody Picture Vocabulary Test (PPVT), measured when the child is 3 or 4.

Table 1: Life Stages and Corresponding Outcomes (Continued)

Stage	Variable	
Early Childhood (Age 5)	Math	Age-standardized scores from the math section of the Peabody Individual Achievement Test (PIAT)
	Reading	Age-standardized scores from the reading recognition section of the Peabody Individual Achievement Test (PIAT)
	Antisocial Behavior	Age-standardized antisocial behavior subscale from the Behavior Problems Index (BPI). Scores are reverse coded so that higher is better.
	Hyperactivity	Age-standardized hyperactivity subscale from the Behavior Problems Index (BPI). Scores are reverse coded so that higher is better.
Middle Childhood (Age 11)	Math	Age-standardized scores from the math section of the Peabody Individual Achievement Test (PIAT)
	Reading	Age-standardized scores from the reading recognition section of the Peabody Individual Achievement Test (PIAT)
	Antisocial Behavior	Age-standardized antisocial behavior subscale from the Behavior Problems Index (BPI). Scores are reverse coded so that higher is better.
	Hyperactivity	Age-standardized hyperactivity subscale from the Behavior Problems Index (BPI). Scores are reverse coded so that higher is better.
Adolescence (Age 19)	High School Graduation Status	A dichotomous variable indicating whether the individual received a high school diploma by age 19. GED earners are not counted as high school graduates.
	Grade Point Average (GPA)	A continuous variable of average grade in the last year of high school. Ranges from 0 to 4.
	Criminal Conviction	A dichotomous variable indicating whether the individual was convicted of any charges other than minor traffic violations by age 19.
	Teen Parent	A dichotomous variable indicating whether the individual reported having a child by age 19.
	Lives Independently from parents	A dichotomous variable indicating whether the individual was living independently from his or her parents at age 19.
	Math	Age-standardized score on a test measuring mathematical ability: math section of the Peabody Individual Achievement Test (PIAT)at age 13 or 14 in the CNLSY and arithmetic reasoning section of the Armed Services Vocational Aptitude Battery (ASVAB), taken between ages 15 and 23, in the NLSY79.
	Reading	Age-standardized score on a test measuring verbal ability: reading recognition section of the Peabody Individual Achievement Test (PIAT) at age 13 or 14 in the CNLSY and word knowledge section in the Armed Services Vocational Aptitude Battery (ASVAB), taken between ages 15 and 23, in the NLSY79.
	Family Income	This continuous variable is the log-transformed measure of the family's income during early adolescence (ideally measured at age 13, 14, 15, or 16).

Table 1: Life Stages and Corresponding Outcomes (Continued)

Stage	Variable	
Adolescence (Age 19) (Continued)	Marijuana Use	This dichotomous variable indicates whether the individual reports having ever used marijuana (CNLSY) or having used marijuana in the past year (NLSY79).
	Other Drug Use	This dichotomous variable indicates whether the individual reports having ever used drugs other than marijuana or amphetamines (CNLSY) or having used drugs other than marijuana in the past year (NLSY79).
	Early Sex	This dichotomous variable indicates whether the individual reports having had sexual intercourse before age 15.
	Suspension	This dichotomous variable indicates whether the individual was ever suspended from school.
	Fighting	This dichotomous variable indicates whether the individual reported getting in a fight at school or work in the past year.
	Hitting	This dichotomous variable indicates whether the individual reported hitting or seriously threatening to hit someone in the past year.
	Damaging Property	This dichotomous variable indicates whether the individual reported intentionally damaging the property of others in the past year.
	Self-Esteem Index	Age-standardized IRT score on the Rosenberg Self-Esteem Scale.
	Religious Service Attendance	This variable measures frequency of religious service attendance on a scale of 0 (none) to 5 (more than once a week).
	Gender Role Attitudes	This continuous variable is the mean of the individual's answers to five questions about how they view women.
	Participation in School Clubs	Dichotomous variable indicating whether the individual participated in clubs in high school such as band, choir, or sports.
Transition to Adulthood (Age 29)	Family income	This continuous variable is the log-transformed measure of the family's income during the year the individual was 29 years old.
	Family income to needs	This continuous variable is the log-transformed measure of the family's income as a percent of the federal poverty during the year the individual was 29 years old.
	College Completion	Dichotomous variable indicating whether the individual obtained a 4-year degree or higher.
	Lives independently from parents	A dichotomous variable indicating whether the individual was living independently from his or her parents at age 29.
Adulthood (Age 40)	Family income	This continuous variable is the log-transformed measure of the family's income during the year the individual was 40 years old.
	Family income to needs	This continuous variable is the log-transformed measure of the family's income as a percent of the federal poverty during the year the individual was 40 years old.

Table 2: Descriptive Statistics for Continuous and Dichotomous SGM Outcomes

Variable	Mean	Std. Dev.	Min	Max
Circumstances at Birth (CNLSY)				
Low Birth Weight	0.08	0.27	0	1
Married Parents at Birth	0.75	0.43	0	1
Family Income at Birth (divided by FPL)	2.79	2.39	0	13
Maternal Age at Birth	26	6	13	47
Maternal Age at First Birth	23	5	13	45
Parenting: cognitive stimulation (standardized)	0	1	-4.8	3.1
Parenting: emotional support (standardized)	0	1	-4.9	2.9
PPVT Age 3/4	0	1	-3.6	5.2
Mother's AFQT	46	29	0	100
Early Childhood (CNLSY)				
PIAT Math Ages 5/6 (standardized)	0	1	-3.1	5.7
PIAT Reading Age 5/6 (standardized)	0	1	-3.2	8.4
Hyperactivity Age 5/6 (standardized)	0	1	-3.2	2.6
Antisocial Age 5/6 (standardized)	0	1	-3.9	2.8
Middle Childhood (CNLSY)				
PIAT Math Age 10/11 (standardized)	0	1	-4.6	3.7
PIAT Reading Age 10/11 (standardized)	0	1	-4.0	3.8
Hyperactivity Age 10/11 (standardized)	0	1	-3.6	2.9
Antisocial Age 10/11 (standardized)	0	1	-4.6	2.9
Adolescence (CNLSY)				
GPA of Last Year of HS	2.92	0.78	0	4
Ever Convicted (Prior to age 19)	0.19	0.39	0	1
Teen Birth	0.13	0.34	0	1
Graduated High School (By age 19)	0.84	0.36	0	1
Lives Independently Age 18/19	0.21	0.41	0	1
Family Income Age 13/14	68,023	47,641	1	156,000
PIAT Math Age 14/15 (standardized)	0	1	-6.2	3.1
PIAT Reading Age 14/15 (standardized)	0	1	-4.7	3.2
Self Esteem Index (standardized)	0	1	-3.5	2.9
Member of Any HS Club	0.66	0.47	0	1
Frequency of religious service attendance	2.94	1.71	0	5
Gender Role Attitudes	2.04	0.50	0	3
Had Sex Before Age 15	0.20	0.40	0	1
Damaged Property of Others	0.13	0.34	0	1
Got in fight at school or work	0.10	0.30	0	1
Hit or Threatened to Hit Someone	0.22	0.41	0	1
Ever Suspended From School	0.15	0.36	0	1
Ever Used Marijuana	0.34	0.47	0	1
Ever Used Drugs (not marijuana or amphetamines)	0.06	0.24	0	1

Table 2: Descriptive Statistics for Continuous and Dichotomous SGM Outcomes (Continued)

Variable	Mean	Std. Dev.	Min	Max
Linking Variables in NLSY79				
Maternal Age at Birth*	26	6	13	48
Maternal Age at First Birth*	22	5	13	44
GPA of Last Year of HS	2.35	1.02	0	4
Ever Convicted (Prior to age 19)	0.09	0.29	0	1
Teen Birth	0.14	0.35	0	1
Graduated High School (By age 19)	0.80	0.40	0	1
Lives Independently Age 18/19	0.43	0.49	0	1
Family Income Age 13/14 (2010$)	60,071	39,219	61	156,000
ASVAB Math Score (standardized)	0	1	-2.7	2.3
ASVAB Reading Score (standardized)	0	1	-3.6	2.2
Self Esteem Index (standardized)	0	1	-3.2	2.9
Member of Any HS Club	0.65	0.48	0	1
Frequency of religious service attendance	3.02	1.68	0	5
Gender Role Attitudes	1.87	0.55	0	3
Had Sex Before Age 15	0.12	0.32	0	1
Damaged Property of Others	0.22	0.42	0	1
Got in fight at school or work	0.29	0.45	0	1
Hit or Threatened to Hit Someone	0.40	0.49	0	1
Ever Suspended From School	0.21	0.41	0	1
Used Marijuana, Past Year	0.48	0.50	0	1
Used Drugs (not marijuana or amphetamines)	0.21	0.40	0	1
Transition to Adulthood (NLSY79)				
4-year college degree by Age 28/29	0.20	0.40	0	1
Live Independently Age 28/29	0.89	0.31	0	1
Family Income Age 28/29 (2010$)	57,279	37,245	61	156,000
Family Income-to-Needs 28/29	3.5	2.6	0	13
Adulthood (NLSY79)				
Family Income Age 40/41 (2010$)	70,274	46,680	7	156,000
Family Income-to-Needs Age 40/41	4.2	3.2	0	13

*Four children were born to mothers age 10 or younger, which we have verified reflects reporting error or other administrative or coding problems in the data.

Table 3: Descriptive Statistics for Categorical SGM Variables

Race (CNLSY)

	Obs	Percent	Cum.
White	4,120	71.2	71.2
Black	819	14.2	85.4
Hispanic	610	10.6	96.0
Other	234	4.1	100

Race (NLSY79)

	Obs	Percent	Cum.
White	4,602	75.5	75.5
Black	740	12.1	87.6
Hispanic	484	7.9	95.5
Other	272	4.5	100

Maternal Education At Birth (CNLSY)

	Obs	Percent	Cum.
Less Than HS	1,478	25.6	25.6
HS	2,239	38.7	64.3
Some College	1,036	17.9	82.2
Bachelor's Degree +	1,030	17.8	100

Maternal Education At Birth (NLSY79)

	Obs	Percent	Cum.
Less Than HS	2,065	33.9	33.9
HS	2,767	45.4	79.2
Some College	673	11.0	90.3
Bachelor's Degree +	593	9.7	100

Table 4: Defining Success at Each Life Stage

Stage	Variables
Success at Early Childhood	Math Score ≥ -1 SD & Reading Score ≥ -1 SD & Antisocial Score ≥ -1 SD & Hyperactivity Score ≥ -1 SD
Success at Middle Childhood	Math Score ≥ -1 SD & Reading Score ≥ -1 SD & Antisocial Score ≥ -1 SD & Hyperactivity Score ≥ -1 SD
Success at Adolescence	Graduated High School (diploma, not GED) & GPA ≥ 2.50 in Last Year of High School & Never Convicted By 19 & Never was a Parent By 19
Success at Transition to Adulthood	Lives Independently from Parents and has either (1) a Family Income To Needs Ratio ≥ 250% or (2) Obtained a College (4-year) Degree
Success at Adulthood	Family Income To Needs Ratio ≥ 300%

Table 5: Imputation in the SGM Dataset

Life Stage	Variable	Final Obs	Regression				Proximity	
			% Imputed	No. Imputed	Corr	R^2	% Imputed	No. Imputed
Core	Race	5,783	-	-	-	-	-	-
	Gender	5,783	-	-	-	-	-	-
Circumstances at Birth	Maternal Age at First Birth	5,783	0	0	-	-	-	-
	Maternal Age	5,783	0.02%	1	0.8	0.80	-	-
	Maternal Education	5,783	1.02%	59	0.37	0.41	18.59%	1075
	Mother's AFQT	5,783	4.63%	268	0.5	0.51	-	-
	Family Structure	5,783	7.40%	428	0.29	0.22	12.14%	702
	Low Birth Weight	5,783	10.43%	603	0.03	0.05	-	-
	Family Income (Percent of FPL)	5,783	11.10%	642	0.44	0.34	20.23%	1170
	Parenting: cognitive stimulation	5,783	20.65%	1194	0.69	0.68	18.50%	1070
	Parenting: emotional support	5,783	21.37%	1236	0.83	0.83	20.15%	1165
	PPVT Age 3/4	5,783	50.96%	2947	0.37	0.35	-	-
Early Childhood (Age 5-6)	Hyperactivity	5,783	15.04%	870	0.14	0.14	11.91%	689
	Antisocial	5,783	15.56%	900	0.29	0.29	12.55%	726
	PIAT Math	5,783	21.30%	1232	0.22	0.20	22.34%	1292
	PIAT Reading	5,783	21.44%	1240	0.36	0.36	23.31%	1348
Middle Childhood (Age 10-11)	Hyperactivity	5,783	19.85%	1148	0.26	0.29	6.42%	371
	Antisocial	5,783	20.89%	1208	0.44	0.43	7.23%	418
	PIAT Reading	5,783	23.97%	1386	0.39	0.39	4.25%	246
	PIAT Math	5,783	24.05%	1391	0.47	0.48	4.18%	242
Adolescence (Age 18-19)	Ever Suspended	5,783	12.35%	714	0.32	0.35	-	-
	Family Income (age 13/14)	5,783	23.41%	1354	0.63	0.99	7.75%	448
	PIAT Reading (age 14/15)	5,783	30.11%	1741	0.61	0.60	36.43%	2107
	PIAT Math (age 14/15)	5,783	30.16%	1744	0.56	0.56	36.50%	2111
	Average HS Grades	5,783	31.00%	1793	0.23	0.22	-	-
	Ever Used Marijuana	5,783	35.17%	2034	0.16	0.12	19.85%	1148
	High School/GED (by 19)	5,783	38.92%	2251	0.33	0.25	-	-
	Self Esteem Index	5,783	40.27%	2329	0.13	0.10	38.92%	2251
	Religious Service Attendance	5,783	40.71%	2354	0.18	0.07	10.69%	618
	Ever Used Hard Drugs	5,783	40.79%	2359	0.15	0.31	10.77%	623
	Ever Convicted (by 19)	5,783	41.31%	2389	0.2	0.19	-	-
	Gender Role Attitudes	5,783	43.92%	2540	0.16	0.16	30.56%	1767
	Living Independently	5,783	44.99%	2602	0.14	0.11	-	-
	Had Teen Birth (by 19)	5,783	45.25%	2617	0.27	0.27	-	-

Table 5: Imputation in the SGM Dataset (Continued)

Life Stage	Variable	Final Obs	% Imputed	Regression No. Imputed	Corr	R^2	Proximity	
Adolescence (Age 18-19) (Continued)	Had Sex Before Age 15	5,783	46.65%	2698	0.25	0.19	-	-
	Got in fight at school or work	5,783	48.85%	2825	0.17	0.19	6.38%	369
	Hit or Threatened to Hit Someone	5,783	48.85%	2825	0.24	0.22	6.35%	367
	Participated in HS Clubs	5,783	63.06%	3647	0.17	0.15	-	-
	Damaged Property of Others	5,783	77.94%	4507	0.17	0.20	12.52%	724
Match Variables in Adult Dataset	Race	6,098	-	-	-	-	-	-
	Gender	6,098	-	-	-	-	-	-
	Had Teen Birth (by 19)	6,098	0.06%	37	0.06	0.07	-	-
	High School/GED	6,098	2.08%	127	0.13	0.09	-	-
	Gender Role Attitudes	6,098	3.16%	193	0.12	0.11	-	-
	Religious Service Attendance	6,098	3.20%	195	0.07	0.02	-	-
	Ever Suspended	6,098	3.90%	238	0.13	0.12	-	-
	Maternal Education	6,098	5.44%	332	0.2	0.21	-	-
	ASVAB Math	6,098	5.64%	344	0.81	0.82	-	-
	ASVAB Reading	6,098	5.64%	344	0.79	0.78	-	-
	Participated in HS Clubs	6,098	6.33%	386	0.2	0.16	-	-
	Maternal Age	6,098	13.25%	808	0.04	0.02	-	-
	Maternal Age at First Birth	6,098	13.25%	808	0.49	0.49	-	-
	Independence	6,098	14.53%	886	0.09	0.08	-	-
	Adol. Income (age 13/14)	6,098	17.84%	1,088	0.75	0.91	63.10%	3848
	Had Sex Before Age 15	6,098	21.63%	1,319	0.15	0.16	-	-
	Hit or Threatened to Hit Someone	6,098	27.66%	1,687	0.08	0.08	-	-
	Got in fight at school or work	6,098	27.68%	1,688	0.29	0.25	-	-
	Ever Used Marijuana	6,098	28.09%	1,713	0.17	0.12	-	-
	Damaged Property of Others	6,098	28.16%	1,717	0.22	0.18	-	-
	Ever Used Hard Drugs	6,098	28.26%	1,723	0.31	0.31	-	-
	GPA Last Year of HS	6,098	33.34%	2,033	0.48	0.48	-	-
	Self Esteem Index	6,098	35.24%	2,149	0.16	0.14	-	-
	Ever Convicted (by 19)	6,098	50.93%	3,106	0.21	0.26	-	-

Table 5: Imputation in the SGM Dataset (Continued)

Life Stage	Variable	Final Obs	% Imputed	Regression No. Imputed	Corr	R^2	Proximity	
Transition to Adulthood (Age 29-30)	Educational Attainment	6,098	2.28%	139	0.6	0.50	-	-
	Live Independently	6,098	3.13%	191	0.1	0.10	4.13%	252
	Family Income-to-needs	6,098	8.00%	488	0.4	0.39	7.23%	441
	Family Income	6,098	8.00%	488	0.74	0.98	7.23%	441
Adulthood (Age 40-41)	Family Income-to-needs	6,098	19.15%	1,168	0.42	0.36	13.73%	837
	Family Income	6,098	19.15%	1,168	0.67	0.99	13.73%	837

Note: The "Corr" column provides the correlation (across non-missing observations) between observed values and values predicted by the regression. The "R^2" column provides the R^2 or pseudo- R^2 values for the models used to impute missing values for each variable. The family income variables have such high R^2 values because they are predicted from, among other variables, income to needs. The low values for the dichotomous variables reflect the well-known problem that linear probability models under-estimate R^2 values (see Greene, 1981, 1983)

Table 6: Correlations between Early Childhood and Middle Childhood Math and Reading, SGM vs. ECLS-K

SGM is above diagonal (in red); ECLS-K is below diagonal (in blue)

	EC Math	EC Reading	MC Math	MC Reading
EC Math	-	.56	.48	.43
EC Reading	.77	-	.45	.52
MC Math	.68	.55	-	.60
MC Reading	.65	.60	.74	-

Table 7: Correlations between Early Childhood and Middle Childhood Behavior

SGM is above diagonal (in red); ECLS-K is below diagonal (in blue)

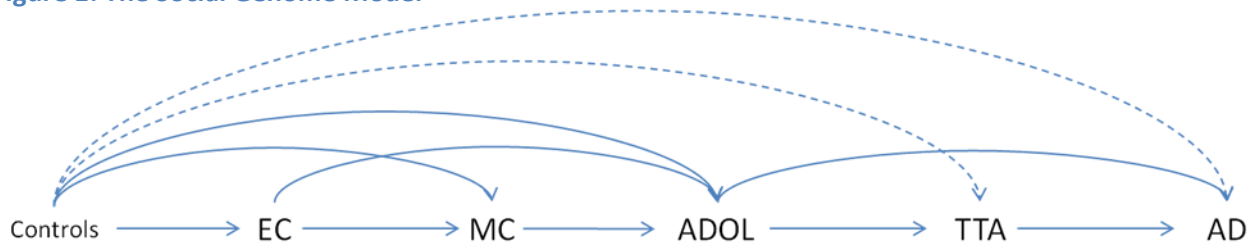
	EC Antisocial	EC Hyperactive	MC Antisocial	MC Hyperactive
EC Antisocial	-	.52	.51	.36
EC Hyperactive	.51	-	.38	.50
MC Antisocial	.42	.31	-	.55
MC Hyperactive	.36	.38	.62	-

Table 8: Family Income Benchmarking

	CPS	PSID	NLSY79	SGM
Overall				
Mean Income to Needs @ Birth	2.7			2.8
Mean Income @ Birth	$53,000			$55,000
Mean Income to Needs @ 40	4.0		4.2	4.4
Mean Income @ 40	$71,000		$70,000	$70,000
Median Income to Needs @ 40	3.3		3.5	3.4
Median Income @ 40	$63,000		$63,000	$58,000
Poor at Birth				
Mean Income to Needs @ 40		3.6		3.0
Mean Income @ 40		$62,000		$49,000
Median Income to Needs @ 40		2.5		2.1
Median Income @ 40		$41,000		$35,000

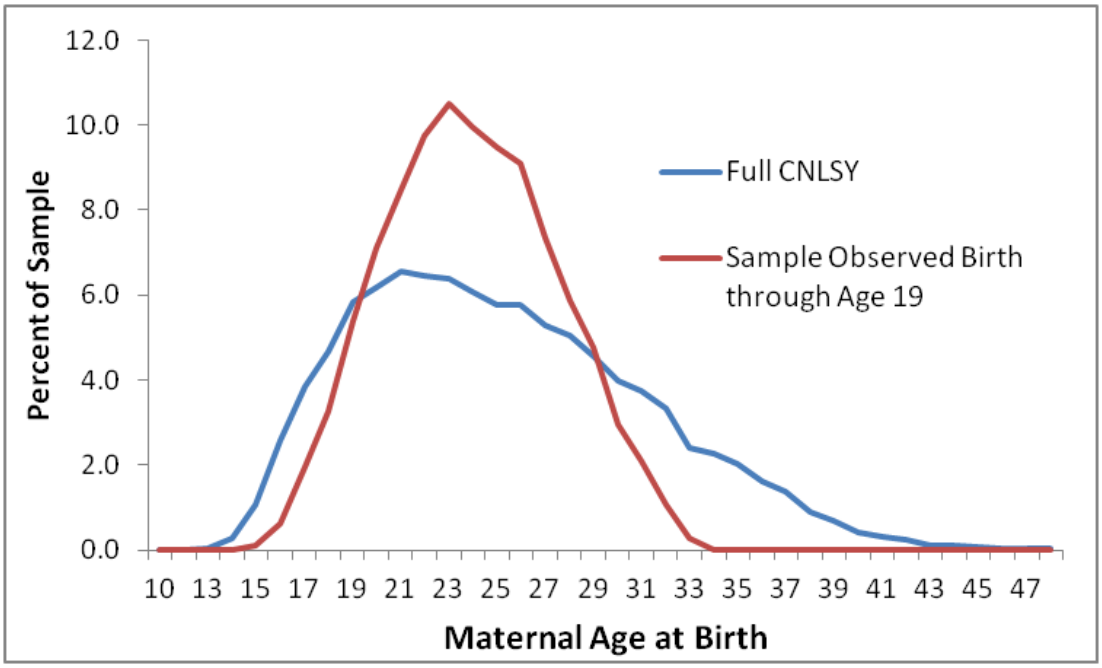
CPS at-birth figures are for 1987 for children under one year old in 1988. CPS at-40 figures are for 2002 for households with a 38-to 42-year-old head in 2003. PSID figures are for adults born between 1966 and 1970, who were age 38-40 in 2006 or 2008. NLSY79 and SGM incomes are topcoded at $156,000; income-to-needs are topcoded at 1300% FPL. All income figures are reported in CPI-U-RS-adjusted 2010 dollars and rounded to the nearest thousand.

Figure 1: The Social Genome Model



Note: Dashed lines indicate that a more limited set of controls is used to predict TTA and AD outcomes.

Figure 2: Distribution of Maternal Age at Birth



"Sample Observed Birth through Age 19" indicates those born 1980 to 1990.

Figure 3: Distribution of Poverty Status in Early Childhood by Race/Ethnicity, SGM vs. ECLS-K

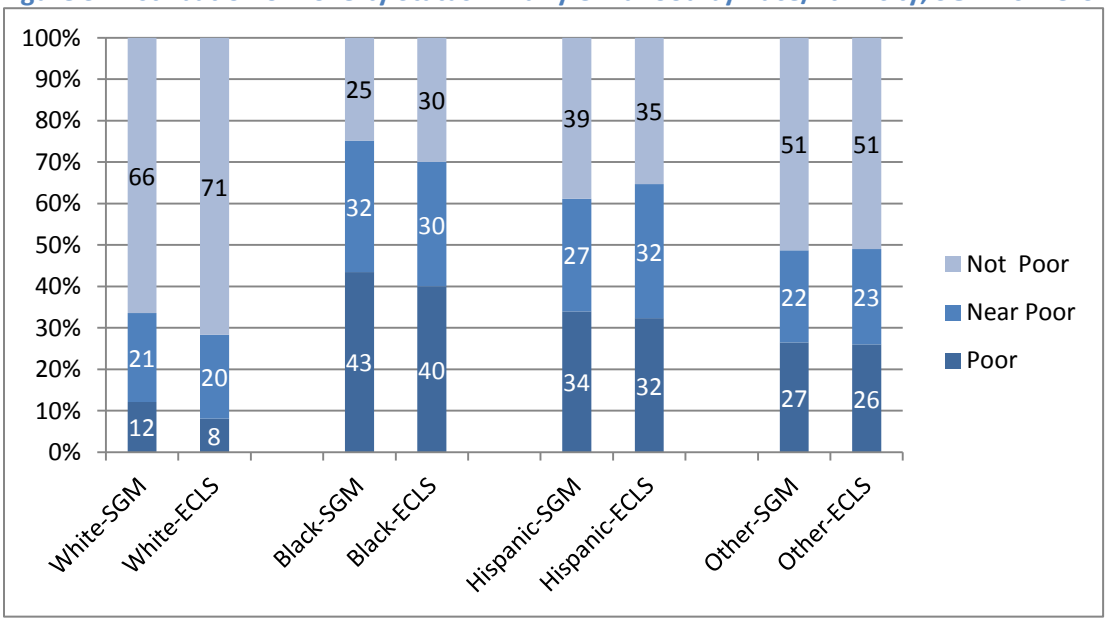


Figure 4: Achievement Gaps by Race/Ethnicity and Poverty Status

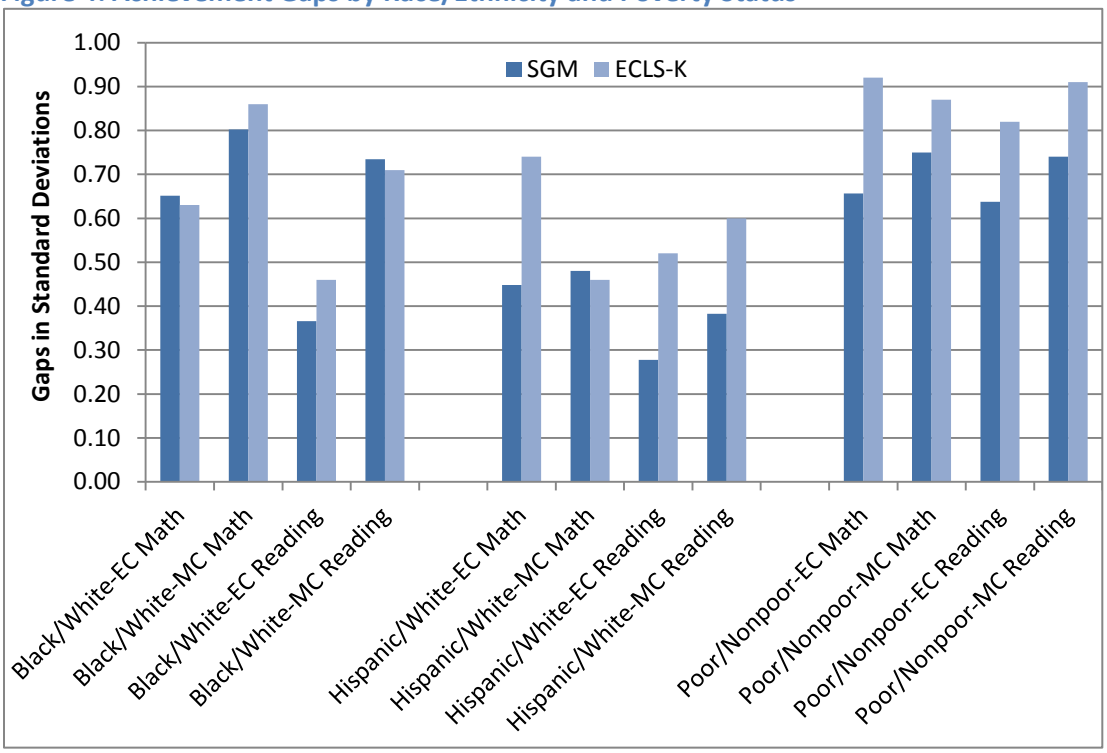
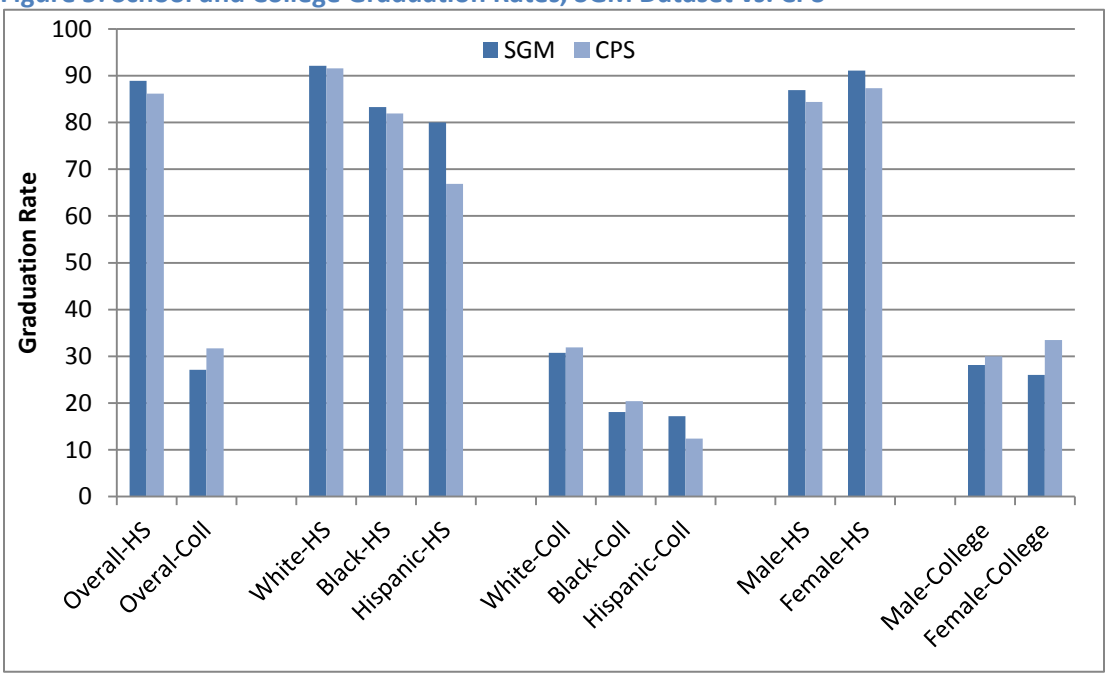


Figure 5: School and College Graduation Rates, SGM Dataset vs. CPS



SGM high school rate is defined here only as percent of sample with a high school diploma or GED by age 19. CPS high school rate is defined as percent of population age 20-24 with a high school diploma or GED in 2005. SGM college rate is defined as percent of sample with a bachelor's degree or higher by age 29. CPS college rate is defined as percent of population age 30-34 that holds bachelor's degree or higher in 2011.