The Social Genome Project at BROOKINGS

Center on Children & Families at BROOKINGS

The Social Genome Project: Mapping Pathways to the Middle Class

April 2013

Overview

The strength of the American political fabric is reinforced by the widely-held notion that each generation will fare better than the one that preceded it. However, the "American dream" is often just that: a dream. More than a fifth of children live in poverty.[1] Additionally, more than 40 percent of children born into the bottom quintile of the income distribution remain at the bottom as they age into adulthood, while a roughly-equal share of children born into the highest quintile remain at the top.[2] Policy makers and policy researchers are thus confronted with the question of what can be done to improve the social mobility of children in economically-disadvantaged families.

The Brookings Institution's Center on Children and Families is constructing a model of social mobility over the life cycle. We believe that this initiative, called the Social Genome Project, will produce a unique new data set and tool for policy analysis.

The model will have a number of advantages:

- It will provide a very explicit and useful life cycle framework for thinking more rigorously about pathways to the middle class.
- It will draw on a variety of existing longitudinal data sources to create a new data set that can then be used to measure a child's chances of success over the life cycle.
- It will create a much-needed complement to an emerging body of research on "what works" based on high-quality evaluations of individual programs because it will enable decision makers to compare and contrast the long-term and indirect effects of different programs to change the life prospects of less-advantaged children and youth. For example, if we know the effects of an investment in early childhood education on a child's chances of being school ready, we will then be able to estimate the longer-term and indirect effects on adult earnings. It could also be used to estimate the effects of multiple interventions on an individual's life prospects.
- It will facilitate benefit-cost analyses of different interventions to help determine where the payoff to a given investment of dollars is likely to have the greatest impact. For example, based on our work so far, we have been able to show that an investment in effective programs that prevent early and unplanned pregnancies would more than pay for themselves, even under very conservative assumptions about the savings to taxpayers.
- It will allow us to examine the distributional implications of different policies, since the model will be based on a detailed representation of the demographic and economic characteristics of the U.S. population.

---

[1] DeNavas-Walt, Carmen, Bernadette C. Proctor, and Jessica C. Smith. *Income, Poverty, and Health Insurance Coverage in the United States: 2009*. United States Census Bureau Current Population Report #P60-238. Washington, D.C.: United States Census Bureau, 2010.

[2] Isaacs, Julia. *Economic Mobility of Families Across Generations*. Policy Report. Washington, DC: The Pew Charitable Trusts, 2008.

- It will serve as a virtual laboratory for conducting new policy experiments and for answering questions about the impact of early success on later success without the costs of mounting a full-scale experiment in the field. Those virtual experiments that seem the most promising could then become candidates for such real-world testing.

While each of these advantages is real, they should not be exaggerated. The model can only be as good as the underlying assumptions and data on which it is based, and care must be taken to subject its implications to a healthy dose of skepticism and to benchmark it against the wisdom and practical knowledge of experts in the field. But even here, its advantage over other approaches is clear; it can be used to test the sensitivity of the results to different assumptions, to identify those knowledge gaps or disagreements that are most critical, and to encourage researchers and practitioners to think more clearly and explicitly about key relationships.

The Social Genome Model

Because it is not possible to capture all aspects of human behavior in a single model, this project's success will also depend on whether the initial architecture of the model is well-adapted to the kinds of policy questions to which it is likely to be applied. For this reason, we have devoted considerable attention to: 1) the specification of our overall policy objectives; 2) the delineation of the key stages of the life cycle, and of the goals at each stage that may contribute to the attainment of those overall objectives; and 3) the solicitation of advice from others on these choices.

The model or framework contains six modules representing different stages of the life cycle from conception to middle age. The data in the model will represent individuals who move through each of these stages with an emphasis on determining how many complete that stage having achieved certain goals such as school readiness by age 5 or high school graduation by age 19.

In conceptualizing our model, we began with an intuitive, simple, and important policy goal: ensuring that as many individuals as possible are middle class by middle age. We will give special attention to those children who begin life in less-advantaged families, although the model will not be restricted to covering the lives of low-income children alone. It will be capable of answering a range of questions about why some individuals – of whatever age or family background – are more successful by middle age than others, and about what policy makers can do to improve people's life prospects. The model will be capable of using many different criteria of success, but in our work so far and in order to make the discussion more concrete, we use a family income of at least 300 percent of poverty by age 40 as one way of quantifying the goal. This is only illustrative of the kinds of outcomes we or other users may want to track. Other possibilities include measuring what proportion of various groups achieves other benchmarks earlier in life. A user interested in a specific question, such as the proportion of African-American children who are reading at grade level by age 10, or the number of poor children who have health insurance, or the number of adolescent boys who have ever been involved with the juvenile justice system, will be able to use the model to answer these and numerous similar questions. These kinds of descriptive uses of the model are already proving to be extremely valuable. Simply tracking the different paths taken by various subgroups of children and the extent to which success does or does not cumulate over the life cycle provides a rich picture of these trajectories.

The model is currently divided into six stages: family formation (conception through childbirth); early childhood (infancy through age 5); middle childhood (ages 5 through 12); adolescence (ages 12 through 19); the transition to adulthood (ages 19 through 29); and adulthood (ages 29 through 40). For each of

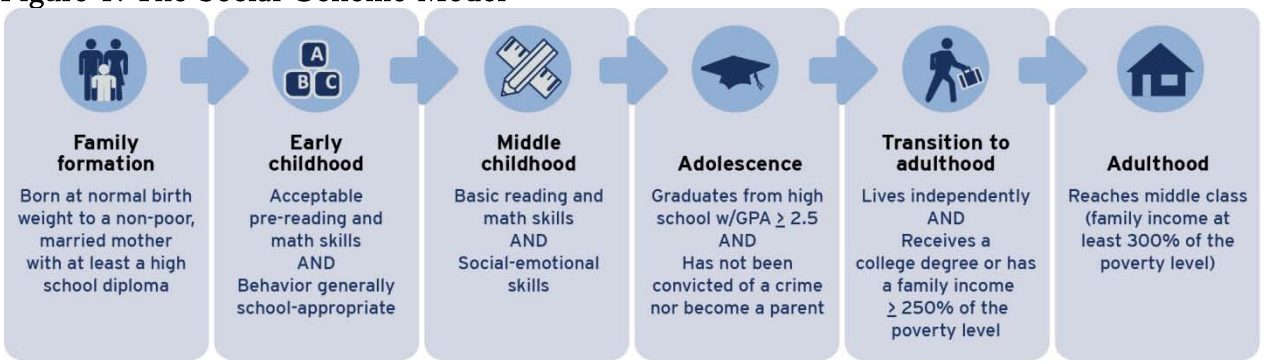
these life stages, we have identified a broad policy goal that, if attained, will help individuals to achieve social mobility or middle class status over the longer run (see Figure 1).

The goals for these six life stages are:

- Family formation: *parental readiness*
- Early childhood: *school readiness*
- Middle childhood: *acquisition of core competencies*
- Adolescence: *college and career readiness*
- Transition to adulthood: *economic self-sufficiency*
- Adulthood: *middle class by middle age*

We have also specified a set of specific sub-goals for each stage that are: a) highly predictive of success at future stages of life, according to empirical data; b) relatively straightforward and easy to explain to the public and to policy makers; and c) measurable using existing data.

Figure 1. The Social Genome Model



Our expectation is that, throughout the modeling process, there will be a tension between our instinct to make the model as sophisticated and comprehensive as possible and our desire to achieve a level of parsimony that makes model-development feasible and policy simulations tractable and transparent. We believe that it will be important for us to remain vigilant in focusing our attention on those factors that are most relevant to our goal of evaluating the efficacy of policies designed to promote social mobility. However, we will also need to avoid the pitfall of limiting the model's usefulness by making it overly simple. Thus, to paraphrase Albert Einstein, the model should be "as simple as possible, but no simpler."

The challenges involved in doing this type of work are many. The data are often imperfect. Identifying causal paths is often daunting. We believe we are well aware of the challenges and pitfalls and we intend to deal with as many as possible and be open and transparent about any remaining sources of uncertainty. One reason we are optimistic is because we have already created a simulation model called FamilyScape, which has given us greater confidence about the feasibility and value of this approach. The Brookings FamilyScape model is a sophisticated simulation model of family formation that simulates the key antecedents of a birth (e.g., sexual activity, contraceptive use, and pregnancy) and many of its most important outcomes, including childbearing among married and unmarried parents and children's chances of being born into poverty. FamilyScape uses a variety of data sources and research findings on the relevant transition probabilities or behaviors and produces rates of pregnancy and childbearing that are well matched to real world outcomes—thereby validating the model's parameters and behavioral assumptions.

The data and modeling limitations that we will undoubtedly encounter serve as another reminder of the importance of maintaining a measure of parsimony as we develop our model. Much of our work will be more descriptive than analytic but in the context of a life cycle framework that allows individual pieces of information or research to fit within a broader and more normative framework. We find the following quote, which was taken from a survey paper on the microsimulation literature, to be a useful admonition in this context:

> "Clear objectives and outcomes are essential to the effectiveness, success, and value of the project…. The potential to add more detail and sophistication to a model will always be a temptation to model developers – and historically this temptation has resulted in many over-ambitious failures."[3]

We are confident that this project is valuable, but we also readily acknowledge that we do not have enough data, manpower, or financial resources to be able to construct a model of social policy that is totally comprehensive. Thus, our project will be guided by the policy questions that we believe to be the most pressing, and by the practical constraints that will inevitably limit the range of policies, behaviors, and outcomes that we are able to model.

Current Status of the Project: April 2013

With generous support from a number of foundations we have made good progress in creating the model, although we have adjusted our approach in a number of ways in response to feedback from our advisers, the availability of funding, and a lot of "learning by doing."

Our accomplishments to date include:

- Creating the Social Genome Model architecture, along with the life stages, and benchmarks of success based on reviews of the literature, discussions with experts, and some new analysis.
- Creating a data set that follows a cohort of children born in the 1980s and 90s to age 40 using actual and imputed data.
- Using our core data set to do preliminary estimates of all the transition probabilities in the Social Genome Model for different subgroups of children and conducting a preliminary analysis of the direct and indirect effects of early success on later success, using structural equation modeling. There are many more questions that we can investigate with the model. For example, we could look at pathways to the middle class for different groups of children, defined by race, gender, circumstances of birth, income quintile, or other variables of interest. We can examine the characteristics of those who are relatively successful and those who are not. We might look at the extent to which falling off the track to success at one or more life stages affects later life stages. The model could help analyze upward and downward mobility across generations and within the life cycle, and the determinants of that mobility. Based on what we find, we strongly suspect that other questions will emerge and that we will want to probe the data for further insights.
- We have done more detailed work on several life stages or modules. This more detailed work includes a well-specified family formation module (conception to birth) which tracks real world outcomes well

[3]Rebecca Cassells, Ann Harding, and Simon Kelly. *Problems and Prospects for Dynamic Microsimulation: A Review and Lessons for APPSIM*. National Centre for Social and Economic Modeling Discussion Paper no. 64. Canberra, Australia: University of Canberra, 2006.

and is now being used to do policy analysis. This additional work also includes an early childhood module and a middle childhood module which are being used to do both descriptive and selected analytic work using a different data set from that employed by the Social Genome Model. We have also partnered with Child Trends to conduct a deeper analysis of adolescence and the transition to adulthood. These supplementary efforts provide a check on the accuracy of the Social Genome Model as well as more detail on what might be driving success in these life stages. With additional funding, we might want to expand this type of module-specific work as it is both valuable in its own right and is also useful for checking the reliability of, or adding data to, the Social Genome Model. Although module-specific work is not a high priority at the current time, we welcome interest from funders who might want to catalyze more work and more policy analysis devoted to a particular life stage.

- We can use the model to simulate the impacts of policy interventions at any point from conception through adulthood on individuals' prospects for achieving later success. Our general strategy has been to link effect sizes drawn from existing randomized controlled trials to the model's ability to translate these effect sizes into the likely long-term effects of such interventions. For now, we would characterize these estimates as preliminary, and somewhat imperfect. However, we are committed to developing the model to improve them, since the only alternative is to do an experiment with an intervention targeted to an existing cohort of very young children and then wait 30 to 40 years to observe what happens to them in comparison to a control group. This is not only very expensive research; it also means missing an opportunity to learn, even if somewhat imperfectly, how to help those children who are already born.
- In addition to producing estimates of the long-term effects of an intervention, the model can also be used to compare different strategies for accomplishing the same goal and for helping to quantify the cost-effectiveness of each. The model can also be used to explore the effects of *multiple* interventions – for example, providing an high quality early childhood program to preschool children, an effective school reform program such as Success for All during elementary school, a small school or career academy experience during high school, and more financial assistance and support during the college years. In these and other such examples, estimating the effects on the proportion of people who are middle class by middle age will require marrying solid research on the effects of a particular program on some immediate outcome to the model's ability to then translate that specific program effect on a particular group into the eventual impact on either the target or a broader group's overall social mobility or chances of being middle class by middle age.
- We and our advisers have worried about both the quality of available data and the assumptions needed to do credible policy analysis with this type of model. The last section of this summary discusses these issues of data and methodology and our response to them in more detail. What we want to emphasize here is that it will take considerable time and effort for our small team to improve the data and the modeling to make them as credible as we and our advisers would like.

Notes on Data and Methods

Although policy makers, journalists, and practitioners have all responded very favorably to our work, some academics have responded more skeptically. Their skepticism is centered on:

1. Well-known problems with identifying parameters in a structural model. Selection effects, omitted variables, and the large number of assumptions needed to specify the equations in the model undermine its credibility.
2. Data and measurement issues, including cohort and period effects, the representativeness of the sample, and problems with matching and imputation.

We take these issues seriously and have addressed them in a variety of ways:

- We have relied on randomized controlled trials or quasi-natural experimental estimates, whenever possible, and given great attention to data issues. We also rely, as much as possible, on external empirical work that is based on a good research design or identification strategy in answering any question. Studies that have used instrumental variables, regression discontinuity methods, differences-in-differences analysis, fixed effects, and the results of randomized trials will be consulted. Estimates based on a credible research design are always to be favored over those based on a more naïve approach. (See especially Angrist and Pischke in the Journal of Economic Perspectives, Spring 2010.) But we also agree with Jim Heckman that it is possible – and often desirable – to combine the two approaches. (See "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy," Journal of Economic Literature, June 2010.) That is, one can use the best estimates from the evaluation literature or from a quasi-experimental study and combine them with a parsimonious structural model in which one doesn't attempt to identify every parameter. The key is to be able to predict a baseline outcome with a set of right-hand side variables and to include a good parameter estimate for the change in outcomes associated with one's policy variable. Note that the emphasis is on getting the policy-induced change in an outcome right and using experimental or quasi-experimental evidence for this purpose. This will always be our preferred approach. Although we have done some modeling within life stages, we have moved away from a focus on within-stage analysis and are making estimating the transition probabilities between life stages in combination with a policy-induced treatment effect the top priority for the future.
- It is worth noting in this connection that the direction of causation in our life cycle model is clear. Outcomes at younger ages affect outcomes at older ages and not the other way around. Thus, concerns about reverse causality are not likely to handicap our effort.
- Most of the relationships underlying the transition probabilities in our life cycle model relate to the effects of academic or social skills on later success – for example, the effects of educational attainment and achievement on earnings, and by extension, on income. The entire literature on the effects of education on later earnings is relevant here. This literature has addressed the kind of selection effects that could be a problem in our analysis. The concern is that more able or motivated students are likely to get more schooling and that their higher earnings may thus reflect the influence of ability and motivation on both schooling outcomes and success in the labor market. However, the more sophisticated literature on the effects of education suggests that those derived from regression models are a good proxy for the causal effect of education on earnings. (See, for example, Rouse and Barrow, Future of Children, Fall 2006. After reviewing the literature, they conclude that "a conventional estimate of the economic value of education is …likely to be quite close to that of the ideal experiment." p. 106). Another reassuring example comes from a paper by Dynarski, Hyman, and Schanzenbach (NBER, Oct. 2011) who show that our type of approach very closely replicates the findings from a long-term follow-up of the children who were in the Tennessee STAR experiment.
- In most cases, a randomized controlled trial is not available for estimating the longer-term effects of an intervention. As noted above, providing such estimates would take decades and be extremely expensive. In such cases, a good-enough but not perfect estimate may be better than no estimate of an impact. In the absence of an informed and empirically-based estimate, policy makers and interested citizens are left with no information on which to base important decisions – such as whether it is better to intervene early or late, whether intervention x has a bigger or smaller long-term effect than intervention y, the likely cumulative effects of several different interventions, or the likely effects of taking a small-scale intervention to scale under certain explicit assumptions about the possible dilution or augmentation of effects as more people are served.

- Integrity and transparency matter. Flaws in data and in parameter estimates should be, and will be, flagged, in papers produced by the project. For example, we recently sent a detailed draft of our preliminary work to all of our academic advisers with extensive details about our data and methodology. We received lots of feedback and are now in a good position to benefit from their advice, improving the data and modeling wherever we can, and being clear where weaknesses still exist.
- Theory, logic, and face validity also matter. If we hypothesize that x affects y, and we get a result that looks ridiculous, we should (and will) recognize that something is wrong.
- Sensitivity analysis, robustness checks, and reporting ranges of effects can also help in providing greater confidence in our findings, and we will employ them where appropriate.