Check for updates

Article

Sociological Methods & Research
2025, Vol. 54(4) 1683–1725
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00491241251347984
journals.sagepub.com/home/smr
Sage

Social Rigidity Across and Within Generations: A Predictive Approach

Haowen Zheng[1] and **Siwei Cheng**[2]

Abstract

How well can individuals' parental background and previous life experiences predict their mid-life socioeconomic status (SES) attainment? This question is central to stratification research, as a strong power of earlier experiences in predicting later-life outcomes signals substantial intra- or intergenerational status persistence, or put simply, social rigidity. Running machine learning models on panel data to predict outcomes that include hourly wage, total income, family income, and occupational status, we find that a large number (around 4,000) of predictors commonly used in the stratification literature improves the prediction of one's life chances in middle to late adulthood by about 10 percent to 50 percent, compared with a null model that uses a simple mean of the outcome variable. The level of predictability depends on the specific outcome being analyzed, with labor market indicators like wages and occupational prestige being more predictable than broader socioeconomic measures such as overall personal and family income. Grouping a comprehensive list of predictors into four unique sets that cover family background, childhood and adolescence development, early labor market experiences, and early adulthood family formation, we find that including income, employment status, and occupational characteristics at early career

[1]Department of Sociology, Cornell University, Ithaca, NY, USA
[2]Department of Sociology, New York University, NY, USA

Corresponding Author:
Haowen Zheng, Department of Sociology, Cornell University, 109 Tower Road, Ithaca, NY 14853, USA.
Email: hz492@cornell.edu

Data Availability Statement included at the end of the article

significantly improves models' prediction accuracy for mid-life SES attainment. We also illustrate the application of the predictive models to examine heterogeneity in predictability by race and gender and identify important variables through this data-driven exercise.

Keywords

social stratification and mobility, life course, machine learning, social rigidity, prediction in social sciences

Introduction

In the stratification literature, the concept of "social rigidity" characterizes the persistence in individuals' socioeconomic status (SES) over the life course and across generations. From descriptive measures such as odds ratios for occupational mobility and intergenerational elasticity for income mobility, to the classic status attainment model established by Blau and Duncan (1967), to modern-day causal mediation analysis that exploits the temporal structure of longitudinal data sets (Brand et al. 2019; VanderWeele and Tchetgen 2017; Zhou 2022a), decades of stratification research has explicated various aspects of social rigidity by analyzing pathways linking individuals' social origin to their mid-life SES. This literature has considered pathways through *intergenerational persistence* (e.g., the effects of parental income and childhood experience) as well as pathways through *intragenerational persistence* (e.g., the effects of early career attainment). Differentiating between different periods of the status attainment process allows researchers to decompose inequality in mid-life SES into distinct mediating mechanisms (e.g., Bloome and Furey 2020; Zhou 2022a).

Current research on the status attainment process typically employs one of two main approaches. The *descriptive approach* focuses on associations between social origin and destination variables, and examines the proportion of variation in the outcome explained by one or more independent variables. The *causal approach* focuses on identifying the causal effects of prior experiences on later status attainment, as well as the mediators that lie along this causal pathway. Traditional approaches rely on structural equation models to unpack causal pathways, often visualized through path diagrams (e.g., Blau and Duncan 1967; Goldberger 1972; Pearl 1998), while modern causal analysis frameworks formalize assumptions through directed acyclic graphs (DAGs) (e.g., Pearl 2012; Robins 2003; VanderWeele and Tchetgen 2017).

This approach unravels the mechanisms of stratification and discriminate among competing explanations for socioeconomic disparities across groups.

We proposes an alternative approach to complement the existing literature by shifting to a *predictive perspective*. Historically rooted in statistics and machine learning, the predictive approach is a relatively recent addition to sociological studies (Brand et al. 2023; Garip 2020; Molina and Garip 2019; Salganik et al. 2020, 2019). A predictive approach to stratification and mobility examines how prior life experiences, including parental background (intergenerational factors) and individuals' own earlier life experiences (intragernational factors), predict later life chance outcomes. In this approach, predictability can be considered as indicator of social rigidity. High predictability of intergenerational factors indicate a persistent influence of parental characteristics on children's outcomes, reflecting a high degree of *intergenerational social rigidity*; high predictability of intragenerational factors reflects a high degree of *intragenerational social rigidity*. The lack of predictability, on the other hand, may stem from uncertainty and unforeseen events over the life course, as well as unmeasured or imperfectly measured factors in the available data set (Lundberg et al. 2024). For example, early life characteristics may be equally influential in shaping later-life outcomes across two societies, but those outcomes may be less predictable in one due to greater earnings volatility in its labor market.

We argue that a predictive approach contributes to the stratification field in three important aspects. First, while previous approaches prioritizes a researcher's point of view, the predictive approach places individuals' perspectives at the forefront. Rather than focusing solely on population-level estimates, such as associations between selected variables or average causal effects, a predictive approach emphasizes individual-level predictions, which offers insights for each unseen data point. Examining predictability helps us understand how individuals' life trajectories might evolve in reality. As such, the predictive approach offers an alternative lens for interpreting the concept of social rigidity. The emergence of strong predictors of life outcomes often signals a persistent continuity of socioeconomic status across an individual's life or between generations. This continuity reflects a higher degree of social rigidity, shaped by the influence of social and institutional factors. For example, strong education-related predictors may suggest rigidity arising from the structure of the schooling system, while the power of early labor market outcomes in predicting later-life outcomes may point to rigidity embedded in labor market dynamics. Conversely, difficulties in predicting life outcomes point towards greater uncertainty of life chances.

Second, adopting a predictive perspective can uncover hidden disparities among different population subgroups, such as race and gender, in terms of predictability and uncertainty. For instance, variations in the predictability of outcomes across demographic subgroups may not align with disparities detected by traditional descriptive measures, such as R^2 used in in-sample analyses. Understanding which population subgroups have the least predictable life outcomes is both substantively important and policy-relevant. If certain subgroups exhibit relatively high predictability, it may indicate that the stratification process for these groups is especially rigid, in the sense that earlier life experiences exert a strong influence on their later-life outcomes. By contrast, if a disadvantaged subgroup's outcomes are particularly difficult to anticipate, this unpredictability may itself contribute to heightened stress and risk in their daily lives and future planning, which merits greater policy attention. Furthermore, the lack of predictability for such subgroups may also reflect how the current understanding of stratification systems may not apply to these groups' life experiences. Hence, this could underscore the importance for the research community to collect richer data—both quantitative and qualitative—to improve understanding of the underlying stratification mechanisms at play for these specific subgroups.

Third, whereas previous approaches primarily rely on historical data, this predictive framework is inherently forward-looking. It focuses on how mobility or immobility can be gauged before outcomes fully materialize, thereby providing policymakers and stakeholders with anticipatory signals rather than merely retrospective evaluations. This shift toward prediction and prevention can guide interventions that are more targeted and timely. Additionally, prediction models developed from one data set can be adapted and applied to others. For example, models trained on older cohorts can be used for younger cohorts. When predictions in new data sets diverge from observed outcomes, this approach provides opportunities to identify gaps or unexpected dynamics, which could lead to valuable insights for refining theories and frameworks.

Beyond the fundamental differences between the predictive approach and previous approaches as outlined above, the practical advantages of a predictive approach should not be overlooked. While previous approaches have focused on explaining within-sample variations, a predictive approach, by emphasizing robust out-of-sample predictability, naturally aligns with machine learning methods designed to mitigate overfitting and model misspecification. This approach enables researchers to leverage a rich toolkit of off-the-shelf algorithms and cross-validation techniques that prioritize reliable predictive performance. Additionally, whereas previous analyses often

focus on a limited set of variables, machine learning methods provide a methodological framework to incorporate a wide array of variables across various life stages, enabling a more wholistic modeling of the long-term status attainment process.

In recent years, machine learning has been increasingly used by social scientists to predict life outcomes that are key to the stratification process (e.g., Breen and Seltzer 2023; Salganik et al. 2020, 2019; Savcisens et al. 2023; Vafa et al. 2022). The literature particularly highlights the importance of combining social scientists' contextual knowledge and insights with machine learning algorithms when defining the question and selecting model inputs (Brand et al. 2023; Garip and Macy 2023; Salganik et al. 2019). Building on this line of work, our study draws on classic theories from inter- and intra-generational mobility research to categorize these variables into four distinct predictor sets corresponding to sequential life stages: (1) parental background, (2) childhood and adolescence development, (3) early labor market experiences, and (4) early adulthood family formation. We begin with models that exclusively focus on family background and progressively incorporate these predictor sets in line with the chronological progression of these life stages. Our goal is to predict mid-life SES achieved between the ages of 40 and 50, a life stage where one's SES tends to stabilize. We also contextualize our findings on the predictability of mid-life outcomes by comparing them with results from previous studies.

The following analyses are divided into two main sections. First, we evaluate social rigidity by analyzing out-of-sample prediction accuracy based on observed variables for mid-life outcomes like hourly wage, income, family income, and occupational status (socioeconomic index [SEI]), while identifying key variables relied upon by the models. We find that complex machine learning models outperform traditional ordinary least squares (OLS) especially when the predictor set exceeds 500 variables, and a benchmark model using the classic Blau and Duncan specification. Second, we reveal variations in predictability and lists of important variables by subgroups defined by race and gender.

Comparing Descriptive, Causal, and Predictive Approaches for Studying Social Stratification and Mobility

As discussed in the "Introduction" section, we propose incorporating a predictive approach into the social stratification and mobility research to

complement the descriptive and causal approaches. To illustrate its unique contributions, Tables 1 and 2 compare the three approaches by five dimensions. The descriptive approach focuses on characterizing and summarizing patterns of stratification and mobility, such as using the proportions in the mobility table cells to characterize absolute class mobility and odds ratios to characterize relative mobility. Similarly, past research has used intergenerational income elasticity and the rank-rank slope for measuring the association between parents' and children's income. The causal approach focuses on identifying mechanisms and quantifying causal relationships that explain why stratification occurs and how specific interventions can change mobility outcomes. In contrast, the predictive approach focuses on accurately predicting future life outcomes in unseen data, asking questions such as: "How well can we predict future life outcomes based on observed prior attainment?" and "Does predictability vary by outcome and subgroup?" Unlike the predominantly backward-looking perspectives of descriptive and causal analyses, the predictive approach takes a forward-looking framework, which captures the perspective of individuals as they navigate their lives and form expectations about their future prospects. This approach also offers an opportunity to evaluate our current understanding of how social stratification works through a data-driven exercise. For example, failures to predict based on a large number of variables selected by an existing theoretical framework can prompt inquiry into why life outcomes are hard to predict, further collection of both qualitative and quantitative data, and readjustments in theories (Garip 2020; Lundberg et al. 2024).

The predictive approach also entails a distinct set of methods, differing from those typically used in descriptive and causal analyses. The descriptive approach primarily uses exploratory tools to elucidate associations between key variables. In contrast, the causal approach relies on (quasi-)experimental data or applies techniques like matching and weighting to adjust for sample selection biases. On the other hand, the predictive approach employs machine learning techniques to flexibly model nonlinear functional forms and complex interactions among multiple predictors, while simultaneously guarding against overfitting through regularization or cross-validation.[1]

Comparing the Predictive Approach With the Classic Status Attainment Model

We next compare our predictive framework with the traditional status attainment model, a key tool for analyzing social stratification processes.

Table 1. Comparison of Descriptive, Causal, and Predictive Approaches for Understanding the Process of Social Stratification and Mobility.

Approach	Descriptive	Causal approach	Predictive approach
Goal	Characterize and summarize patterns.	Identify mechanisms and causal relationships.	Accurately predict or classify outcomes in unseen data.
Focus	Documenting *what is happening* in social rigidity (e.g., the association between parents' and children's socioeconomic position).	Understanding whether and *why* interventions can change mobility outcomes.	Determining *how predictable* socioeconomic outcomes are, based on known and observed factors.
Keyquestions	What are the levels, trends, and variations in the determinants of socioeconomic attainment over the life course and across generations?	What factors (e.g., parental background, education, location, and discrimination) cause or mitigate the unequal distribution of socioeconomic attainment?	How well can we predict future life outcomes based on prior attainment? Does predictability vary by outcomes and subgroups?
Methods	Exploratory tools like summary statistics, cross-tabulations in occupational mobility tables, odds ratios, and intergenerational income elasticity.	Structural equation models, experimental or quasi-experimental methods (e.g., RCTs and instrumental variables), matching and weighting methods that correct for sample selection.	Machine learning models (e.g., regression trees, random forests, and neural networks), or regression with a focus on prediction and accommodation of a large number of variables.
Key strengths	Easy to interpret; provides a general understanding of the associations of pre-identified socioeconomic attainment indicators.	Provides insight into underlying processes, causal mechanisms, and insights for policy interventions under clear assumptions.	Prioritizes individual life experiences; evaluates the existing theories and knowledge framework; emphasizes a forward-looking research perspective.

Abbreviation: RCTs = randomized controlled trials.

Originating from Blau and Duncan (1967), this approach explicates the relationships between socioeconomic variables using a path diagram. A classic example is Blau and Duncan's path analysis, which incorporates five variables: father's education, father's occupation, respondent's education, first job, and occupation in 1962. These variables are interconnected by lines indicating correlation (e.g., a curved line with arrowheads at both ends between father's education and occupation) and directed arrows signifying influences from upstream to downstream variables (e.g., an arrow from father's education to respondent's education). In this paradigm, the main quantities of interest include total and partial effects of the upstream variable on the obtained SES, the strength of which indicates how strongly SES is passed down across and within generations. These estimands are usually estimated as path coefficients through structural equation models, a system of interrelated equations. In model specification, the status attainment model framework often relies on assumptions about the causal ordering of a limited number of variables and employs linear models with no explicit interactions between variables.

Besides direct and indirect effects, another quantity of interest in the status attainment model is the in-sample R^2, which is used as a metric of model fitness, as well as to measure the amount of variation in status attainment that is accounted for by upstream variables. For example, Blau and Duncan (1967:165-172) were able to account for 33 percent of the variance in (current) occupational (prestige) attainment of a national sample of men in 1960s, by using their father's occupational prestige, respondent's own educational attainment and first job. Similarly under the broad umbrella of the status attainment framework, the Wisconsin school of social psychological models showcase the usefulness of social psychological concepts in stratification processes by adding variables like mental ability and influence from significant others into models that further explains away variation in levels of educational and occupational attainment (Sewell et al. 1969). Though informative, the limitation of using R^2 terms stems from the fact that they are merely in-sample statistics, which naturally increase as one adds more variables into the model and can lead to model overfitting. Therefore, in-sample R^2 terms are limited in helping us learn about the predictability in other data sets or situations beyond the specific conditions of the original data.

Our predictive approach diverges from the classic status attainment models by emphasizing prediction accuracy as a measure of social rigidity. Rather than focusing on the statistical association among a few selected variables, we reformulate the question of social rigidity as a prediction task: given an individual's information is measured at various life stage points, how well

Table 2. Comparison of Methods: Practical Differences.

Analytic approach	Analytic tools (examples)	Quantities of interest	Number of variables	Model specification
(a) Classic status attainment model	Path diagram; structural equation models; mobility table methods	Total effects; partial effects; direct and indirect effects; and in-sample R^2	Limited	Typically linear models; no explicit interaction
(b) Conventional mediation analysis	Directed acyclic graph; linear regression models	Natural direct effect; natural indirect effect	Limited	Typically linear models with two-way interactions
(c) Machine-learning-based mediation analysis	Directed acyclic graph; machine learning methods	Natural direct effect; natural indirect effect	Moderate	Flexible non-linear models with complex interactions
(d) Our predictive framework	Multi-stage directed acyclic graph of wholistic sets of variables; machine learning methods	Prediction accuracy (out-of-sample R^2); predictive group-specific distributions	Large number of variables; variable sets spanning across multiple periods	Flexible non-linear models with complex interactions

can we predict what will happen to them at a future time point? Our key quantities of interest include out-of-sample R^2 values and predictive distributions for specific groups. While status attainment models often rely on a limited set of variables and use parametric methods like linear regressions, our predictive framework offers greater flexibility in both the number of variables and the model specification. Contemporary machine learning methods, adept at handling a large array of variables through feature selection and cross-validation, effectively mitigate the risk of overfitting. This adaptability is crucial for exploring the intricate intra- and inter-generational dynamics of stratification, as highlighted by existing literature emphasizing the complex interplay of numerous variables throughout the life course (Athey and Imbens 2016; Brand et al. 2021). For instance, to accurately reflect one's social origin, a broader range of variables beyond just father's education and occupation is essential. Additionally, our approach relaxes assumptions about functional forms to account for the potentially complex interrelations among multiple predictors of socioeconomic outcomes. This allows for a more nuanced and comprehensive understanding of the factors influencing socioeconomic trajectories.

Comparing With Traditional and Contemporary Mediation Analysis

Researchers have also used various mediation analysis to characterize the transmission of SES across generations and over the life course. We categorize mediation analytical frameworks into two more detailed categories: the conventional mediation analysis, and machine-learning-based medication analysis (for a more comprehensive review of causal mediation analysis, see Brand et al. 2023).

In general, mediation analysis frameworks primarily aim to understand the direct and indirect effects of a variable on an outcome by identifying the pathways of influence. These frameworks are particularly useful in unpacking the process of status attainment, such as how family background influences one's mid-career SES through schooling and employment trajectories, which can span across both inter- and intra-generational contexts (e.g., Bloome and Furey 2020; Brand et al. 2019). Traditional mediation analysis often relies on linear models supported by DAGs. However, more recent advancements have introduced more versatile machine learning methods. These modern approaches yield model-free estimates and can handle a moderate number of variables (e.g., Brand et al. 2023; Pearl 2012; VanderWeele and Tchetgen 2017). Meanwhile, researchers exploring multiple, time-ordered mediators in stratification or mobility processes often turn to sequential

mediation analysis. This method, like our predictive approach, is adept at modeling multi-stage processes, such as those aligned with life stages, and frequently employs machine learning models for estimation (e.g., Imai and Yamamoto 2013; Vansteelandt and Sjolander 2016; Zhou 2022b). The key distinction between our predictive framework and sequential mediation analysis lies in their respective focuses. While sequential mediation analysis focuses on stage-specific causal effects of particular treatments or interventions, our predictive framework focuses on the concept of social rigidity, as reflected in the predictive power (measured as the out-of-sample R^2) of different sets of earlier-life-stage predictors.

In summary, by focusing on how much previous life stages can predict mid-life SES, our study joins a growing number of studies that aim to understand the predictability of life trajectories (Badolato et al. 2023; Salganik et al. 2019; Savcisens et al. 2023). For example, in the fragile families challenge, social scientists aimed to predict specific outcome variables at the age of 15. Despite the availability of data encompassing thousands of variables collected to enhance the understanding of these families' lives, researchers were not able to make accurate out-of-sample predictions. Efforts to measure the predictability of life outcomes not only inform our understanding of social rigidity from a social stratification point of view, questions on how to improve prediction accuracy could also lead to further developments in theory and methods. Building on a rich literature in inter- and intra-generational social mobility, our study is the first to investigate the predictability of important life course socioeconomic outcomes at midlife.

Data and Methods

Data and Variables

The main source of our data is the National Longitudinal Survey of Youth 1979 (NLSY79). Directed by the U.S. Bureau of Labor Statistics, The NLSY79 is a representative sample of 12,686 individuals born between 1957 and 1964, who resided in the United States at the start of the survey (Bureau of Labor Statistics, 2023). The initial interview took place in 1979 when the respondents were between 14 and 22 years old. Subsequent interviews were conducted annually from 1979 to 1994, and then every 2 years. We use data from 1979 to 2016, which covers the life experiences of the participants from their early teenage years to the age of 50. We exclude the military sample and a subset that discontinued participation in 1990.

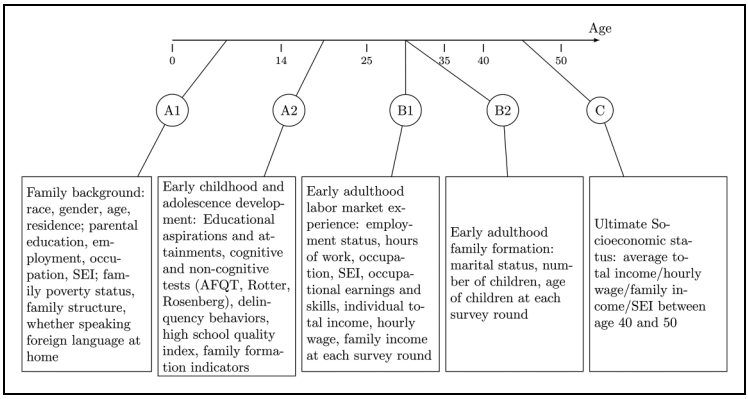

Figure 1. Chronological overview of prediction and outcome sets with representative variables.

Other data sources include the Occupational Information Network (O*NET), from which we construct occupational skill variables, and the 1990 U.S. Census, from which we construct occupational earnings. As a primary source of occupational information, the O*NET is a comprehensive database of worker attributes and job characteristics. Information is collected from and developed by incumbent workers and occupational analysts (National Center for O*NET Development, 2022).

As shown in Figure 1, we assemble five sets of variables that capture an individual's life course experiences: family background (A1), early childhood and adolescence development (A2), early labor market experience (B1), early adulthood family formation (B2), and mid-career socioeconomic attainment (C). We group the variables into distinct sets based on age. Figure 1 illustrates the representative variables from each set for clarity.

Set A1: Family background variables describe one's basic demographic characteristics and parental background. Basic demographic variables include age, gender, race, residence (in the United States or not, in the South or not, and in rural areas or not). Parental background variables include parents' educational attainment (in years), parents' labor market attachment (whether they are working for pay, whether they are working more than 35 hours per week), parents' occupation and SEI (Duncan SEI score), family poverty status, family structure (a dummy indicator for two parent families and a categorical indicator for all family types as specified in the survey[2]), sibship size, and whether the family speaks foreign languages

at home. Most of these variables measure respondents' family background at or before age 14 through retrospective questions about their circumstances at that time.

Set A2: The second set of predictors also describes one's pre-labor market characteristics, measured before age 25. First, it includes a set of cognitive and psychological factors commonly used in the literature, such as the Armed Forces Qualification Test (AFQT 1981), the Rotter Locus of Control Score (1979), and the Rosenberg Self-Esteem Scale (1980). This set also includes measurements on respondents' juvenile delinquency activities in 1980, and whether they are ever charged by the police for engaging in illegal activities. We include respondents' educational expectations (i.e., the respondent expects to be in school in the next 5 years) and aspirations (highest grade one would like to obtain), employment and occupational expectations in 5 years, and a friend's educational aspirations. Second, this set includes educational attainment variables (high school degree, college degree, and advanced degree), and for those with college degree, their college major. Third, we exploit a rich set of high school contextual variables, such as the number of books in the school library, school-level dropout rate, ethnic/racial composition of students and faculty, and so on. Lastly, we include a set of early family formation indicators to capture their family status before they formally enter the labor market. These variables include a respondent's marriage status and whether the respondent has a child by age 18.

Set B1: The third set of predictors includes measures of respondents' early labor market experience from age 25 to 35. Specifically, the variables include employment status, hours worked in a week, occupations (in detailed Census categories), occupational skills, occupational earnings, occupational SEI, and individual income variables. We construct occupational skill variables that are commonly used in the literature, including verbal, interactive, computer, science-engineering, creative, quantitative, supervising, and analytical skills. We extract these occupational skills following the standard practice in the literature by conducting factor analysis (Cheng et al. 2019; Liu and Grusky 2013). For those who report a non-missing occupation, we also merge occupational earnings—measured as the average income within an occupation as reported in the 1990 Census—into the data set. For occupational socioeconomic status, we used gender-specific Hauser-Warren SEI scores and Fredricks Hauser-Warren-equivalent scores updated to the 1980 and 2002 occupational codes (Frederick and Hauser 2010; Stevens and Cho 1985; Warren et al. 1998). Income variables include log total income, log hourly wage, and log total family income measured at each survey round.

Set B2: The last set of predictors describes the formation of the family between the ages 25 and 35. These variables include marital status, the number of children, and the age of the children if the respondent has children at each survey round. We include this set as recent research suggests that events occurring in other domains of life, such as the family domain, could impact one's attainment in the labor market (Budig and England 2001; Cheng 2014; Killewald 2013).

Set C: We construct C set as the outcome set to measure one's mid-life SES attainment. These variables include the average log total income, average log hourly wage, average log family income, and average occupational SEI between ages 40 and 50. We choose this specific age range for constructing outcomes of interest as past research shows that for most individuals, this period is characterized by peak or stabilized income levels, and it also shows the strongest correlation with one's lifetime earnings (Haider and Solon 2006; Mazumder 2005). In the main analysis, we focus on predicting income and SEI. We also present results from an additional analysis, where we predict respondents' probability of employment between ages 40 and 50.

To address panel data attrition and missing data, we restrict our analysis sample in several ways. First, to mitigate the impact of sample attrition on the sparsity of life course data, we include respondents in the analysis sample only if they provide a minimum of five data points in the B set (out of eight survey rounds, which dropped 9 percent of the sample) and two data points in the C set (out of six survey rounds, which dropped another 16 percent of the sample). To maintain a maximum number of observations, we conduct missing data imputation. For predictor sets on family background and early childhood and adolescent experiences, we change the missing value in the original variable to the minimum value in the variable and create a flag variable for each change.[3] For predictors that cover one's early adulthood labor market and family formation experiences, we fill in missing value with non-missing value from the closest survey year between ages 25 and 35. All income variables are inflation-adjusted to 2018 US dollars before log transformation. For those with 0 income, we added 1 before log transformation. The analytic sample contains 7,463 respondents who have complete data for the variables of interest after missing data imputation and satisfy the sample restriction criteria above. The data contains 4,253 predictors across all life stages.[4]

Machine Learning Models and Evaluation Metric in Main Analysis

In the analysis, we use various machine learning models to perform the life course outcomes prediction tasks. We start with using set A1 as the prediction

set, and then sequentially add set A2, set B1, and set B2 into the model to predict the outcome variables in set C.

For the main analysis of all four outcomes, prediction sets, and machine learning models, we partition the analysis sample into two subsets for simplicity. The training set contains 70 percent of the data while the test set contains 30 percent.[5] We then run models on the training set and report R^2 value from the test set as a common evaluation metric for comparison of fit across models. To evaluate performance from the machine learning models across prediction sets, we use three benchmarks. The first is a benchmark model that uses the mean of the outcome variable from the training set as the predicted value (hereafter referred to as the "null model"). This benchmark null model will be reflected in our calculation of the test R^2 as an evaluation metric for predictive performance. The second benchmark model is a simple ordinary linear model (OLS), using the same prediction set as the machine learning model. The last benchmark model is a traditional Blau and Duncan model (hereafter referred to as the "B&D"). This model is specified based on the original status attainment model, which is an ordinary linear regression modeling using father's educational attainment, father's occupational status, respondent's own educational attainment, and respondent's first occupational status (Blau and Duncan 1967).

We rely on two kinds of machine learning methods to perform life-course prediction tasks. The first kind is shrinkage regressions, namely the Ridge and Lasso regression methods. Ridge regressions are similar to the OLS, except that the coefficients are estimated under a shrinkage penalty λ that shrinks some coefficients toward zero. With this shrinkage penalty parameter, Ridge regression's advantage over OLS is rooted in the bias-variance trade-off. At $\lambda = 0$, the Ridge regression is equivalent to the OLS, where the variance is high but there is no bias. As λ increases, shrinking the Ridge coefficients leads to a substantial reduction in the predictions' variances but at the expense of an increase in bias. Ridge regression outperforms OLS in situations where there is high variance, where a small change in the training data set can lead to a large change in the least-squares estimates, especially when the number of variables is as large as or even bigger than the number of observations. While the Ridge shrinks the coefficients towards but never reaches zero, the Lasso regression has a variable selection property by shrinking some of the coefficients to zero. This means that the Lasso regression is likely to produce simpler and more interpretable models compared with the Ridge, if the outcome is better characterized as a function of a small set of predictors. However, in terms of prediction accuracy, they can return similar results due to qualitatively similar model behaviors (i.e., as λ increases, the variance decreases and the

bias increases). The shrinkage methods are helpful in our case, as one of our research goals is to expand the number of predictors in each life stage. We use cross-validation for selecting the tuning parameter λ.

The second set of models consists of tree-based ensemble methods, including random forest, gradient boosting, and Bayesian additive regression trees (BART). These models aggregate multiple regression trees, which stratify or segment the predictor space into smaller and simpler regions in order to achieve the greatest possible reduction in the residual sum of squares. In each region, the regression tree model then uses the mean of the outcome variable for the training observations as prediction values for the test observations that fall within the same regions. Thus, tree-based methods are likely to outperform models based on linear regression in cases where the relationship between the outcome and predictor variables are highly non-linear and complex. However, individual trees can be sensitive to minor variations in the data, leading to significant changes in the final estimated tree model. To overcome this disadvantage, tree-based ensemble methods combine a large number of "weak learner" trees into a more robust and potentially powerful single model. Here, we focus on three commonly used ensemble methods: the random forest (which averages regression trees that only pick a random sample of predictors through bootstrap sampling), boosting (growing trees sequentially by fitting new trees to the residual of the current fit), and BART (the model grows trees successively but perturbs each tree to avoid local minima). We explore the tuning parameters through cross-validation.[6] We compared these models with a SuperLearner ensemble, which selects from various algorithms, including train set means, OLS regression, Ridge, Lasso, tree-based models, and neural networks with variable screeners. The SuperLearner consistently favored tree-based models and regularized regressions, yielding test R^2 results similar to tree-based ensembles (see Appendix A2, Supplemental Material for details).[7]

In summary, we will benchmark the performance of machine learning models against that of a "null model," a traditional OLS model with the same set of predictors as the machine learning model, and a B&D OLS model. Machine learning models have the advantage over OLS by accommodating a broader array of predictors, while also capturing non-linear relationships and interactions among these predictors. Techniques like Ridge and Lasso regression, known for shrinkage or variable selection, become particularly advantageous as the number of predictors grows. Additionally, tree-based methods prove valuable for their capacity to handle complex model specifications, including flexible interaction terms between multiple predictors.

Following common practice, we report the test R^2 as our metric to evaluate predictive performance:

$$R^2_{test} = 1 - \frac{\Sigma_{i \in test}(y_i - \hat{y}_i)^2}{\Sigma_{i \in test}(y_i - \overline{y}_{training})^2}, \quad (1)$$

where $\overline{y}_{training}$ is the mean of the outcome variable in the training data set obtained from the null model. The R^2_{test} is one of the commonly used evaluation metric in the prediction literature (Salganik et al. 2019). In our case, this measure is particularly helpful for comparison across outcomes, as it does not depend on the variance of the outcome variables.

Important Variables

We use SHapley Additive exPlanations (SHAP) values to generate lists of the top 20 important variables from gradient boosting models, using various prediction sets to predict log hourly wage (Lundberg 2017).[8] We focus on log hourly wage as it has the highest predictability among the outcomes. SHAP assigns each variable a contribution value for individual predictions, enabling a more nuanced understanding of feature importance compared to traditional methods, which often rely on aggregate measures like Gini importance or regression coefficients. It uses fair allocation results from cooperative game theory to allocate credit for a model's output among its input features, providing consistent measurements regardless of feature correlations (Lundberg 2017; Lundberg et al. 2020). We incorporate 5-fold train/test split cross-validation in the calculations for robustness and reducing the risk of overfitting.[9] By splitting the data into five subsets, we train the model on four folds and test it on the fifth, repeating the process across all folds to get an average performance estimate. Additionally, we produce lists of the top 20 important variables for predicting key intermediate variables, including AFQT and average log hourly wage in the B set.

Predictability and Important Variables by Race and Gender

The last section examines predictability by race-gender subgroups. We train a gradient boosting model using all predictors across life stages on a combined sample of white and black respondents that include both women and men. The model is then evaluated on test data by both race and gender to estimate test R^2 values on white women, white men, black women, and black men. We evaluate predictability and key variables across race and gender, drawing on

prior literature on intersectional inequality, which demonstrates that race and gender jointly shape labor market outcomes in stratification processes (Browne and Misra 2003; Cheng 2016; McCall 2005). As in the previous section, we use 5-fold cross-validation for the train/test split and apply 500 bootstraps before the train/test split to calculate the 95 percent confidence intervals for test R^2. To identify important variables by race and gender, we train models on group-specific subsamples and generate importance rankings for each group, using SHAP values.

Results

Predictability of Life Course Outcomes

How well can mid-life SES be predicted? Which life stage predicts one's SES attainment well? Do results vary by prediction model? Table 1 presents our evaluation metric, R^2_{test}, for two selected outcomes, log hourly wage and log total income. We make three main observations from Table 3.

First, the predictability of life course outcomes depends on which outcome we examine. Prediction accuracy is significantly higher for the average log hourly wage between age 40 and 50 than for total income for the same age range. For example, the best R^2_{test} statistics in Panel 1, from models that use tree-based ensemble methods and sets A1, A2, and B1, is 0.53. However, the best R^2_{test} in Panel 2, which comes from the same model and set combination, is only about 60 percent (0.33). Among the best performing machine learning methods, the R^2_{test} from using set A1 only (e.g., Panel 1, 0.25 from BART) exceeds the R^2_{test} from using sets A1 and A2 combined (Panel 2, 0.15 from BART). Overall, the predictability for total income ranges from about 30 percent to 60 percent of that for hourly wage. Second, predictability varies across prediction sets and improves as additional sets are introduced in life course order.[10] Across both panels for the two outcomes, a consistent pattern emerges: Set B1, which captures early adulthood labor market experiences, stands out as the most effective predictors. This aligns with previous research indicating that predictors temporally closer to the outcome often exhibit greater predictive power (Cheng et al. 2014; Kim et al. 2018).[11] However, it is notable that Set B1 is more predictive than Set B2 (early adulthood family formation experiences).

The third main observation arises from comparing horizontally across models. Consistent with expectations, the simple OLS model performs similarly to the more complex machine learning models when prediction sets are relatively small. For example, when the prediction set only includes set A1,

Table 3. Comparison of Test R^2 for Log Hourly Wage and Log Total Income Across Models.

	Methods					
	OLS	Ridge	Lasso	Random forest	Gradient boosting	BART
Panel 1: Log hourly wage						
A1	0.22	0.23	0.23	0.24	0.24	0.25
+ A2	0.31	0.34	0.35	0.33	0.35	0.36
+ B1	−1.21	0.47	0.51	0.52	0.53	0.53
+ B2	−1.34	0.47	0.51	0.52	0.53	0.53
B&D	0.24	–	–	–	–	–
Panel 2: Log total income						
A1	0.07	0.08	0.08	0.07	0.09	0.08
+ A2	0.08	0.13	0.13	0.13	0.15	0.15
+ B1	−1.84	0.30	0.31	0.31	0.32	0.33
+ B2	−1.87	0.31	0.32	0.32	0.33	0.33
B&D	0.10	–	–	–	–	–

Note: OLS = ordinary least squares; BART = Bayesian additive regression trees; B&D = Blau and Duncan; SEI = socioeconomic status.
Prediction accuracy is measured by test set R^2. R^2_{test} from the OLS B&D model for family income is 0.16, and 0.30 for SEI.
Data Source: National Longitudinal Studies of Youths 1979 cohort, 1979–2016.

which contain fewer than 500 features, its R^2_{test} is only marginally lower than the more complex machine learning methods. However, when the number of features dramatically increase to over 4,000 after we add sets A2 and B, the OLS's predictions cannot match shrinkage methods like ridge and lasso. This is expected as shrinkage methods are designed to improve prediction accuracy by giving more weights to predictors that contribute the most to the model, while penalizing or "shrinking" the coefficients of less important variables to reduce overfitting and improve model performance. In addition, the ensemble methods, which allow complex non-linear relationships and interactions, perform only marginally better than shrinkage methods. Comparing across linear shrinkage methods and flexible tree-based methods suggests that the prediction task is improved primarily through expanding the number of variables in the prediction sets, rather than incorporating complex non-linear relationships.[12]

We next contextualize these results by discussing their relations to two comparison cases. The first comparison case is the classic B&D model of

status attainment. Here, we compare our R^2_{test} based on our machine learning models with the R^2_{test} that would result from estimating model that includes the variables used by the B&D model. While the B&D model provides a reasonable baseline (e.g., R^2 of 0.24 for log hourly wage and 0.10 for log total income), incorporating additional predictors and using regularized or machine learning methods substantially improve predictive accuracy. For instance, the inclusion of 4,000+ predictors (models with "+B2" data) boosts test R^2 of log hourly wage to 0.53 for tree-based methods like random forests and gradient boosting, and 0.33 for log total income. Results are similar for SEI and log family income, with the B&D model achieving test R^2 values of 0.30 and 0.16, respectively (Table 3 notes). These levels are comparable to those of more complex machine learning models when limited to A set predictors but fall significantly below the performance of models that incorporate all 4,000+ predictors (Figure 2). These results highlight that while the traditional model captures key aspects of attainment, high-dimensional data combined with machine learning methods offer significantly better predictive performance, underscoring the value of broader data collection and advanced methods in modeling midlife outcomes.

The second comparison case is the fragile families challenge we mentioned earlier, which reports the highest out-of-sample R^2 at ~0.2 for predicting material hardship and GPA at age 15, and this metric drops to around 0.05 for other outcomes (Salganik et al. 2020). The fact that prior life experience can predict as high as half of the variations of SES at mid-life implies substantial persistence of socioeconomic indicators both over the life course and across generations. However, when interpreting our results in comparison with other studies, one should note the important differences in research design. First, we examine outcomes at a different life stage. Mid-life socioeconomic outcomes may be easier to be predicted based on sociological theories, compared with outcomes at adolescence, the target life stage in the fragile families challenge. In contrast to younger ages, individuals' positions in mid-life tend to exhibit greater structure and stability. This phase of life is influenced by various social institutions that contribute to shaping one's trajectory, including acquiring educational qualifications, forming family bonds through assortative mating, and embarking on distinct career paths in the labor market. Second, we examine different target populations. Our study is based on a cohort-specific survey that best characterizes the life trajectory of the late Baby Boomers,[13] while the fragile families challenge draws from data that over-sample births to unmarried mothers, resulting in the inclusion of a large number of black, Hispanic, and low-income families (Reichman et al. 2001). Therefore, difference in the out-of-sample R^2 can also stem

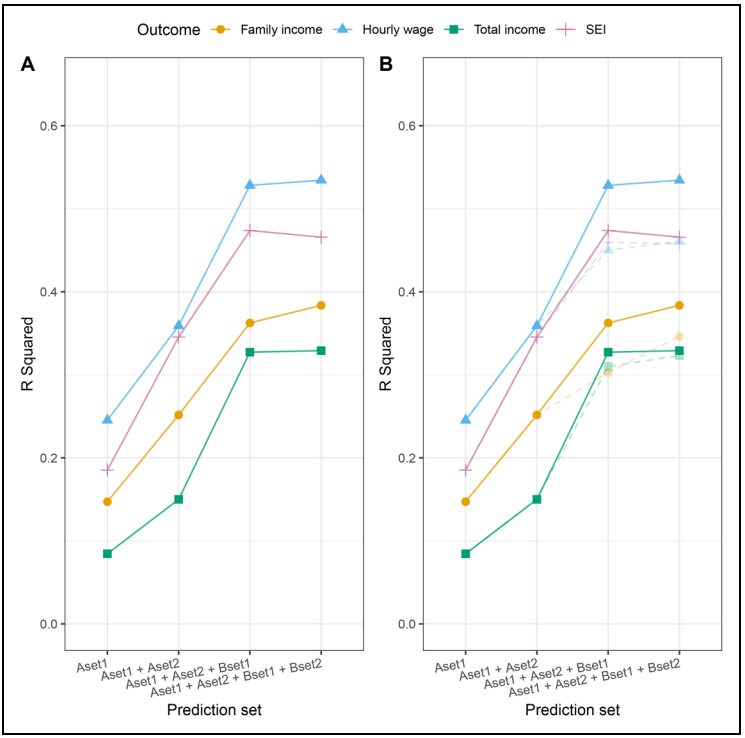

Figure 2. Predictability across outcomes and sets. *Note:* The left panel shows results before dropping "lag" variables, and the right panel shows results after dropping "lag variables" by adding transparent dots and dashed lines. For outcomes that include hourly wage, total income, and family income, the "lag variables" include all individual level wage and income variables in set B1 and thus subsequent modeling. For predicting SEI, the "lag variables" include all "SEI" variables in set B1 and subsequent modeling. SEI = socioeconomic index.

from the inherent variation in predictability of life course outcomes across social groups. Lastly, we examine different outcomes. Focusing on social rigidity, we select traditional indicators from the stratification literature. As our results imply, the level of predictability depends on which outcome we examine. This is likely the case when comparing across studies. Specifically, income and occupational outcomes might be easier to predict than outcomes such as student GPA, material hardship, and layoff.

What do these prediction models tell us about the level of social rigidity for this cohort of people overall? We consider a stratification system highly rigid if mid-life outcomes can be accurately predicted using variables commonly referenced in the literature. The scatterplots in Figure 2 present a more comprehensive illustration across all outcomes and prediction sets, particularly highlighting results by BART, one of the best performing models across outcomes. The data reveals that predicting total income and family income poses the biggest challenge, while hourly wage and occupational "SEI" are relatively more predictable. Overall, the predictability varies widely, ranging from as low as below 0.1, to as high as just over 0.5.

Consistent with Table 3, Figure 2 also suggests that adding set B1, which includes early adulthood labor market experience, into the prediction set, leads to the largest increase in R^2_{test}.[14] Consistent with past studies, we find that these "lag variables" hold the largest predictive power (Kim et al. 2018). But how much of the relatively high predictability of set B1 is attributable to these "lag variables?" In Panel B of Figure 2, we drop those variables from the prediction sets and redo the prediction tasks. The transparent dots and dashed lines present the updated results after dropping those "lag variables." We see that dropping those variables does decrease the R^2_{test} for models that include set B1. When using all variable sets to predict log hourly wage, dropping all individual level income variables leads to a 15 percent decrease in R^2_{test} (from .53 to .46), and for log family income, it leads to a 12 percent decrease (from .38 to .34). For predicting mid-career occupational SEI, dropping all SEI variables from set B1 does not decrease the R^2_{test} as dramatically. This suggests that the majority of variation in mid-life occupational status is predominantly accounted for by other variables, such as individual income levels and occupational characteristics (e.g. occupational earnings and occupational skills). Interestingly, omitting all individual income variables from the models has minimal impact on the predictions for mid-life log total income. This could be attributed to the inherent difficulty in predicting total income, regardless of whether these "lag variables" are included in the analysis. Notably, Figure 2 suggests that there is considerable level of social fluidity as the R^2_{test} ranges from 0.3 to 0.53 across the four outcomes, when using all the prediction sets combined together to do the prediction task, even after including all the "lag variables."[15]

A Closer Look at Important Variables

We use SHapley Additive exPlanations (SHAP) values to identify the most influential variables for predicting mid-life SES attainment, using the gradient

boosting method. The average absolute SHAP values indicate the extent to which each feature influences the model's prediction, by quantifying how much it shifts the output from the baseline value (the mean value from training data) to the specific model output. Figure 3 presents the top 20 most important variables across the four prediction sets for predicting log hourly wage.[16] Note that SHAP values reflect the correlations captured by the predictive ML models. They should not be interpreted as causal estimates, as they do not indicate how the model's predictions would change under interventions on specific variables. Establishing causality requires constructing causal models, which involves making explicit causal assumptions. Instead, our purpose here is to present features most strongly associated with the outcome in the context of using a large number of life stage indicators for the prediction task. This exercise provides substantive insights into patterns that may warrant further exploration, even if they are not causal.

The top left panel shows results from a model that only includes the prediction set on family background. Out of 99 predictors, the variables that contribute most significantly to the reduction in the outcome's variance include gender, age, parents' educational attainment and SEI, family poverty status, race, sibship size, region, and so on.[17] The results align with findings from the stratification literature on intergenerational social mobility, as the variables ranked most critical for accurately predicting log hourly wage are those commonly identified as core concepts. Interestingly, this data-driven approach highlights the significance of maternal socioeconomic characteristics, such as highest grade completed, occupational status, full-time employment status, and whether she held a professional occupation. Additionally, the prominence of age likely reflects the survey design, where family background variables, such as family poverty status, were only collected for younger respondents in the initial wave.[18]

The top-right panel shows results from the model incorporating all pre-labor-market predictors. When both A1 and A2 predictors are included, the AFQT score, which measures cognitive abilities, emerges as the most influential variable. Other highly ranked predictors include gender, age, college degree, educational aspirations, the Rosenberg self-esteem score, race, and family background factors. While occupational expectations and college major (particularly nursing) appear lower on the list, high school context variables were not among the most critical predictors for accurately forecasting midlife log hourly wage. The model relies most heavily on AFQT, gender, age, and college degree to make predictions.

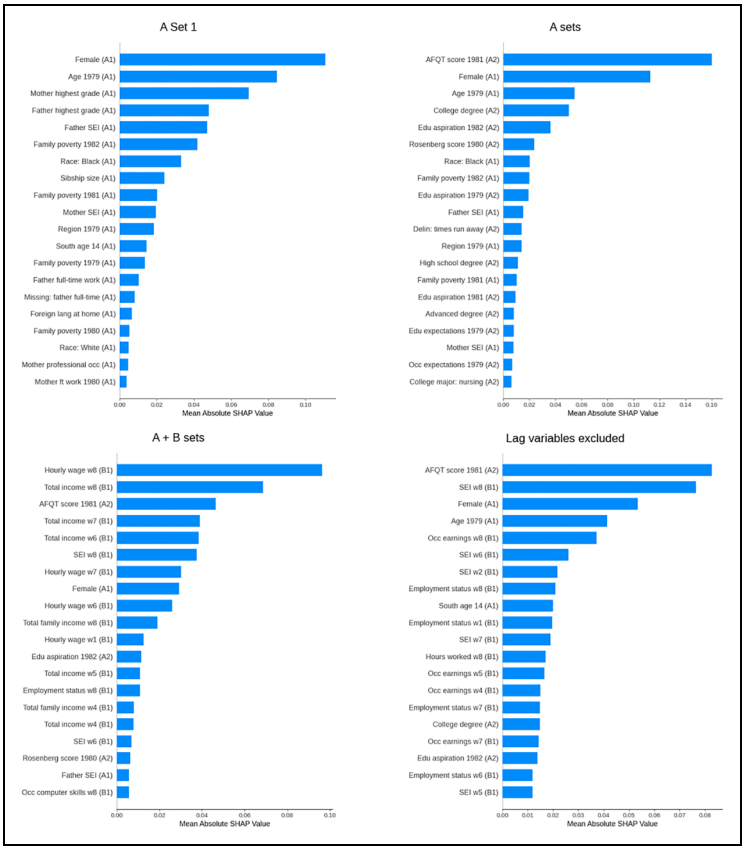

Figure 3. Important variables for predicting log hourly wage by sets. *Note:* (1) SHapley Additive exPlanations (SHAP) values are calculated using XGBoost with 5-fold train-test split cross-validation. (2) For the sake of having shorter Y labels, some labels are cut short. "lang" represents "language"; "occ" "occupations"; "ft" "full-time"; "edu" "education"; and "delin" "delinquncy."

The bottom-left panel presents results from a model that incorporates all predictors across life-course stages.[19] Consistent with prior findings that identify set B1 (labor market experiences) as the most critical predictor set for midlife SES, the model predominantly selects B1 variables as the most influential among over 4,000 features. Furthermore, the results suggest that income and wage variables from later waves within the B1 age range (25–35) are more predictive of midlife SES attainment than occupational

indicators, such as occupation dummies, SEI scores, occupational earnings, and occupational skills, aligning with past research (Kim et al. 2018).

However, the analysis also shows that not all variables close to midlife are equally predictive of midlife labor market outcomes. For example, family formation variables are not selected among the most important variables. In addition, the results reaffirm the importance of the AFQT score, which the model relies on even more heavily than some income and wage variables in early waves of B1. When lagged B1 income and wage variables are excluded, employment status during early adulthood emerges as a significant predictor, and occupational SEI appears more predictive than occupational earnings, likely due to SEI's incorporation of both education and earnings.

As previously mentioned, although it is valuable to pinpoint important predictors, such as wage and income variables in B1 and the AFQT cognitive score in A2, we must be cautious in interpreting these as evidence of strong causal effects of the variables themselves. In fact, it is plausible that these variables themselves are determined by upstream variables, such as family background in A1. To get a sense of the endogeneity of these important intermediate variables, we conduct a similar exercise aimed at predicting these variables. Table 4 presents the results. As anticipated, a considerable amount of variation in these important intermediate variables can be predicted by variables that measure one's family background and other pre-labor market characteristics, with the R^2_{test} ranging from 0.24 to 0.45.

Figure 4 presents a similar analysis using SHAP values to identify important upstream predictors for AFQT and the B set1 average log wage. Consistent with findings from previous studies (Fischer et al. 1996), AFQT scores are strongly associated with a range of predictors, including race, which reflects historical and structural inequalities, alongside family background characteristics such as parents' educational attainment, occupational status, and family structure. For the important variables identified when the prediction task focuses on average log hourly wage, as shown in the right panel, several socio-demographic factors emerge as strong predictors, including age and gender. Additionally, AFQT, region, college degree attainment, and delinquency indicators are highlighted as significant predictors of early adulthood wages. The prediction also heavily relies on education-related variables, such as educational degrees, aspirations and expectations, and high school contexts, including total enrollment numbers and teacher characteristics. While this article does not delve into specific variables on this list, our examination of key variables in the prediction models provides direction for future research to identify important mediators in the life course stratification process.

Table 4. Predicting Intermediate Variables Using Upstream Predictors.

ImpVar	Prediction set	Test R^2
AFQT (A2)	A1	0.43
Average log hourly wage (B1)	A1 + A2	0.35
Average log total income (B1)	A1 + A2	0.24
Average log family income (B1)	A1 + A2	0.26
Average SEI (B1)	A1 + A2	0.45

Note: AFQT = Armed Forces Qualification Test; SEI = socioeconomic index.
Prediction accuracy is measured by test set R^2.
Data Source: National Longitudinal Studies of Youths 1979 cohort, 1979–2016.

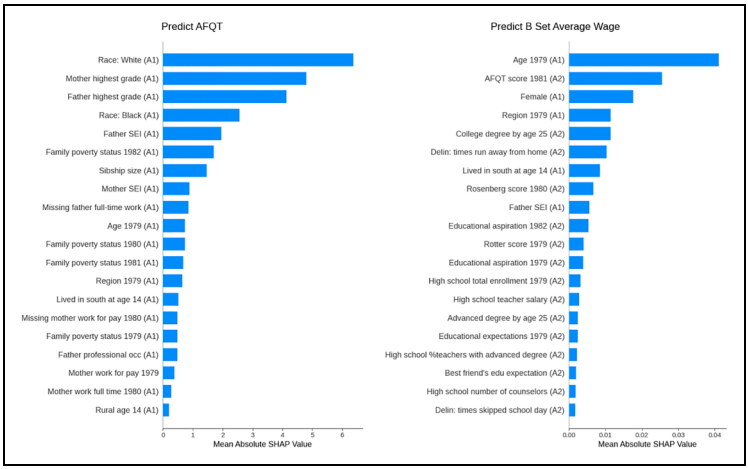

Figure 4. Important variables for predicting intermediate variables. *Note:* SHapley Additive exPlanations (SHAP) values are calculated using XGBoost with 5-fold train-test split cross-validation.

Intersectional Predictability

Figure 5 presents predictability of midlife SES by race and gender. The bar graph illustrates the distribution of test R^2 across different subgroups and outcomes. The test R^2 values represent how well the model, trained on the combined sample, performs in predicting the different outcomes for each subgroup (defined by race and gender). A higher test R^2 value indicates

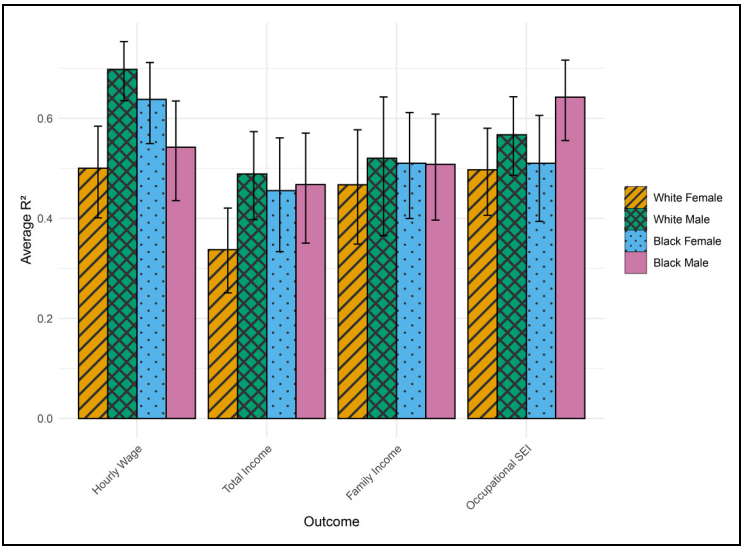

Figure 5. Average Test R^2 by outcome and subgroup. *Note:* R^2_{test} are calculated using gradient boosting with 5-fold train-test split cross-validation and 95 percent confidence intervals are based on 500 bootstraps.

greater accuracy in the model's predictions for the outcome within that specific subgroup on unseen test data.

Results vary by the outcome of interest. For log hourly wage, the model performs best for white males (test $R^2 = 0.70$), followed by black females (test $R^2 = 0.64$), indicating strong predictive accuracy for these groups. In contrast, white females (test $R^2 = 0.50$) and black males (test $R^2 = 0.54$) show lower test R^2 values, suggesting that the model is less effective in predicting hourly wages for these two subgroups. Figure 6 provides a closer examination of predicted values versus actual values in the test set. It shows that the prediction model, trained on joint samples, is most likely to yield predictions resembling the distribution of white men in unseen data. The bias for other groups primarily results from the model over-predicting log hourly wages for extremely low actual values. In the case of white women, the bias is also driven by the model's inability to accurately predict higher wages for some data points. These results align with previous labor market studies that highlight stronger labor market attachment and stable wages among white males and black females overall and at critical

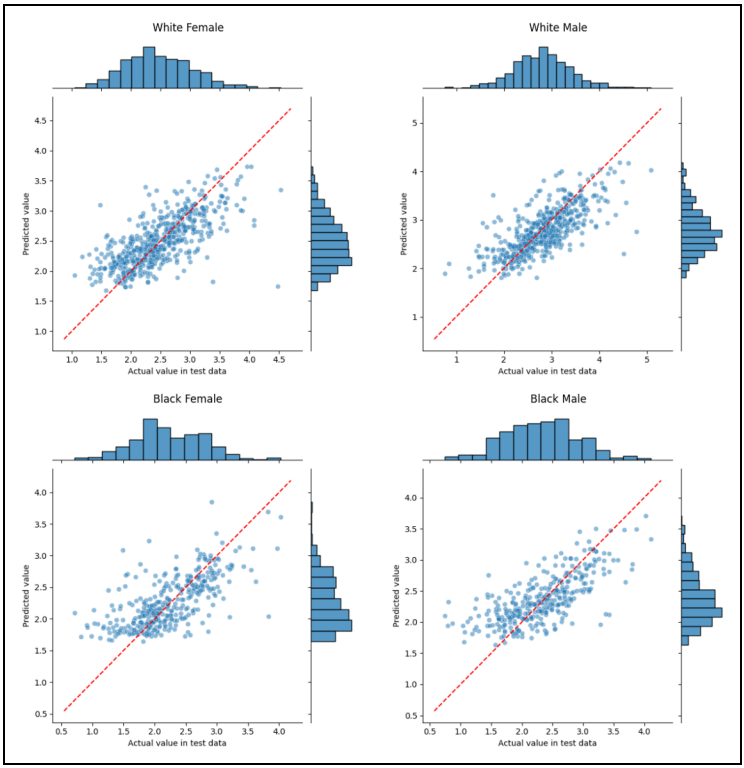

Figure 6. Predicted log hourly wage and actual test values, by race and gender. *Note:* Results are calculated using gradient boosting.

life transitions such as parenthood (Cheng 2016; Glauber 2007; Sweeney 2002). Research has consistently found that white males tend to have more stable and continuous employment patterns, often due to greater access to opportunities and resources. On the other hand, black females typically demonstrate higher labor market participation that are historically stable, which is often driven by the need to support their households and navigate economic challenges, and a greater emphasis on and social values attached to employment stability and engagement in the workforce compared with their white counterparts (Goldin 1977).

When it comes to predicting midelife SEI, the model exhibits the highest predictability for black males with a test R^2 value of 0.64, followed by white males (test $R^2 = 0.57$). The higher predictability among black males is likely

not a sign of labor market advantage but rather due to the clustering of SEI values in the lower range, meaning both the real and predicted values are low (see more details in Appendix A5, Supplemental Material). Black females (test $R^2 = 0.51$) and white females (test $R^2 = 0.50$) show more modest results, but the predictive performance is still relatively strong across all groups. This is consistent with results in Figure 2, which shows predictabilty for SEI ranks high among all four major outcomes. For log total income and log family income, the differences between groups become less pronounced. However, white females still rank the lowest across both outcomes, with the model showing weaker predictive accuracy for this subgroup. This may reflect differences in income sources or other unobserved factors that the model does not capture as effectively for white females compared to others.

The disparities in predictive performance across these groups may arise from differences in the underlying relationships between the predictors and outcomes. For instance, a regular model, which does not account for race or gender, might overlook the nuanced ways in which these factors intersect to affect outcomes. This could lead to biased predictions, particularly for minority groups, where unique socioeconomic patterns may not be adequately captured. By failing to incorporate these variations, a standard model risks underestimating the complexity of socioeconomic disparities, which may result in lower predictive accuracy for minority groups.

To further investigate the gender and racial variations in key variables identified through this data-driven approach, Figure 7 presents the results of SHAP value analysis by subgroup for predicting log hourly wage. The lists of important variables are produced by running gradient boosting models separately on samples by race and gender. The main text focuses on results that exclude lagged income and wage variables from the B1 set, as including these lagged variables produces similar patterns across subgroups, with the top 20 variables dominated by these lagged predictors (see more details in Appendix A6, Supplemental Material).

The analysis reveals both similarities and distinct patterns by race and gender. A commonality across all groups is the prominence of occupational SEI—an index combining occupational level education and occupational earnings—as the strongest predictor, particularly in the timeline closest to midlife. Additionally, both AFQT and age emerge as significant predictors for all groups.

In terms of between-group differences, first, employment status and hours worked are consistently selected as important predictors for all groups except white males. Interestingly, employment status, in particular, is repeatedly

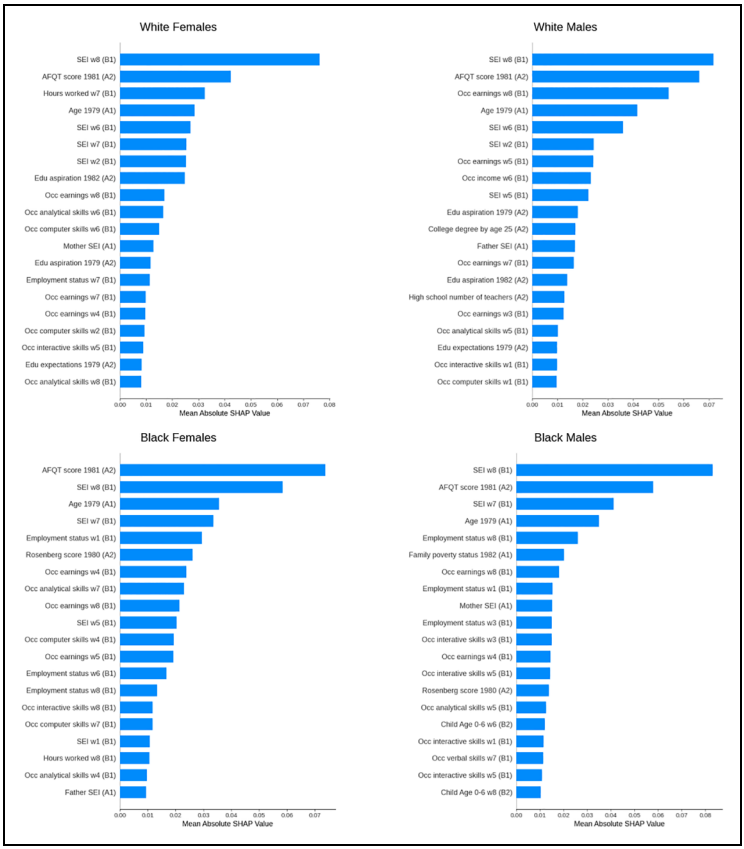

Figure 7. Important variables by race and gender, without lag variables. *Note:* Results are calculated using LightGBM with 5-fold train-test split cross-validation.

selected as a key predictor for black females and males across different waves in early adulthood. Second, for black males, the model places considerable weight on family structure in early adulthood, a pattern that does not emerge for other groups. Third, occupational skills are ranked higher as a predictor for minority groups compared to white males. Finally, the model relies more heavily on education-related variables for whites than for blacks, highlighting potential differences in the factors influencing early adulthood wages across these subgroups. Together, these findings emphasize the importance of

viewing predictability as a group-specific, heterogeneous aspect of the stratification process.

Additional Analysis: Predicting the Probability of Employment

In the main analysis, we focused on predicting mid-life income and SEI attainments. This means that the analytical sample was restricted to respondents with at least one non-missing income and occupation value in survey waves between ages 40 and 50. This criterion may lead to the exclusion of respondents who are consistently unemployed. To the extent that the life chances of employed individuals tend to be less volatile and hence more predictable, the results from main findings can be considered as an upper bound for our ability to predict the cohort's life chances. To supplement the main analysis, we conduct an additional analysis on predicting the probability of employment during age 40 and 50, where we expand the main analytical sample to also include those who are consistently unemployed and create an indicator for the probability of employment. For each respondent, we calculate the probability of employment as a continuous variable, defined as the number of survey waves in which they were employed divided by the total number of survey waves with non-missing data.[20] We then conduct similar analysis using both the OLS and more complex machine learning models. We evaluate the model performance using the same metric, R^2_{test}.

Table 5 presents results from this supplementary analysis. Overall, the results are consistent with our main findings with regard to patterns across prediction sets and models. While only using family background and early childhood and adolescent variables leads to a R^2_{test} of below 0.1, adding labor market variables significantly improves prediction accuracy by almost doubling the R^2_{test}. While the OLS model is widely off when the feature space expands significantly after including set B1, the ML model R^2_{test} hovers right below 0.3. However, the general level of predictability is lower than all the previous outcomes. This suggests that employment status among the general population is harder to predict than income or occupational status among the employed workforce.

Discussion

How persistent is SES across generations and over the life course? How predictable is mid-life SES? This study reframes "social rigidity" within a predictive framework. Leveraging machine learning techniques and measuring predictability by out-of-sample R^2, we find that the

Table 5. Test R^2 for Probability of Employment Across Models.

	Methods					
	OLS	Ridge	Lasso	Random forest	Gradient boosting	BART
A1	0.07	0.08	0.08	0.07	0.09	0.09
+ A2	0.07	0.12	0.12	0.11	0.13	0.13
+ B1	−2.14	0.26	0.28	0.27	0.28	0.28
+ B2	−2.18	0.26	0.28	0.28	0.29	0.29

Note: OLS = ordinary least squares; BART = Bayesian additive regression trees.
Prediction accuracy is measured by test set R^2.
Data Source: National Longitudinal Studies of Youths 1979 cohort, 1979–2016.

predictability for mid-life SES ranges from 0.1 to 0.5, depending on the outcome and predictor set. Our best-performing models improve upon a baseline model (using training set mean) by up to 50 percentage points. Our study contributes to a growing strand of social science research that treats prediction as a means of understanding empirical regularities—particularly in the contexts of intergenerational and intragenerational social mobility (Badolato et al. 2023; Breen and Seltzer 2024; Lundberg et al. 2024; Salganik et al. 2019; Savcisens et al. 2023). Compared to findings from the fragile families challenge—where the highest out-of-sample R^2 for predicting material hardship at age 15 is ∼0.2—our results suggest that mid-life SES is considerably more predictable, with the highest R^2 reaching around 0.5.

Compared with previous descriptive and causal approaches to studying social stratification and mobility, the predictive perspective offers three complementary advantages: it prioritizes individual life experiences, offers an opportunity to evaluate the existing knowledge framework, and emphasizes a forward-looking research perspective. It places individuals' life experiences at the forefront by incorporating a wide range of variables and enabling a flexible, data-driven framework that more closely reflects the complexity of real-life trajectories in life course progressions. Practically, the predictive approach employs advanced machine learning methods with flexible model specification while simultaneously mitigating overfitting. It also identifies the key variables that the model relies on most heavily for its predictions. While these variables may not necessarily represent causal relationships, understanding their significance can still provide valuable insights into patterns and factors shaping life outcomes.

We first apply this predictive framework to gauge the overall level of predictability for a general understanding of social rigidity. We apply state-of-the-art machine learning models to the NLSY79 data, a widely recognized longitudinal dataset in the stratification literature. We select variables commonly used in previous stratification research and group them into four distinct prediction sets that characterize both intergenerational and intragenerational aspects of social rigidity: family background, childhood and adolescence experiences, early adulthood labor market experiences, and early adulthood family formation processes. Predictability increases as we sequentially incorporate prediction sets that align with the life course's temporal sequence. Using a set of predictors that include as comprehensive as 4,000 variables, the level of predictability depends on the specific outcome being analyzed, with labor market indicators like wages and occupational prestige being more predictable than broader socioeconomic measures such as overall personal and family income. While not addressed in this study, future research can usefully explore the reasons behind these variations by outcome and investigate whether "compound outcomes" with multiple components are more difficult to predict.

Relatedly, recent research has shown that life course outcomes are generally difficult to predict, even with rich longitudinal data and advanced machine learning techniques (Badolato et al. 2023; Breen and Seltzer 2024; Lundberg et al. 2024; Salganik et al. 2020). While our models achieve somewhat higher predictive accuracy—outperforming results from prior studies such as the fragile families challenge, where the highest R^2 was around 0.2 (see the "Results" section for more detailed discussions)—they still reveal considerable uncertainty in forecasting mid-life SES. On the one hand, our ability to predict a substantial portion of the variation in mid-life outcomes may reflect a consistent empirical pattern linking earlier-life experiences on later outcomes—one shaped by the complex interplay of structural constraints, individual agency, and simply, luck, as individuals navigate social institutions as they progress through the life course and accumulate life experiences. On the other hand, the inability of prediction could also raise many meaningful questions, such as how social scientists should interpret this unpredictability, and whether better theories and measures of social factors could improve forecasts of key life course outcomes (Garip and Macy 2023; Lundberg et al. 2024). For example, our results show that much of the improvement in prediction accuracy stems from the expansion of the predictor sets rather than allowing for complex interactions between variables, underscoring the importance of incorporating a diverse range of measurements for social concepts and mechanisms in the initial stages of

data collection. Finally, the often limited accuracy of social outcome predictions cautions against social institutions or organizations relying on predictive algorithms to make critical decisions about life chances, such as eligibility for social benefits, medical insurance, bank loans, or immigration outcomes, at an era of growing reliance on AI systems (Eubanks 2018).

To explore group variations, we also apply this predictive framework to examine group differences in predictability by race and gender. Our analysis shows that when predicting hourly wages, the predictive model trained on the combined sample performs best for white men, followed by black women, black men, and white women. Analyzing individual predictions in the test data reveals a tendency for the model to overpredict outcomes for minority group members. However, these differences are substantially reduced when examining other outcomes. Interestingly, we find high predictability for black men's occupational status. This result, however, reflects a form of social rigidity that exacerbates their disadvantage, as the high predictability arises from both predicted and actual values being consistently low. Analyzing the importance of variables across subsamples by race and gender reveals distinct patterns in the variables the model relies on most for accurate predictions. These differences in predictability may partly reflect variation in how strongly observed variables relate to outcomes across groups. However, they may also point to missing or mismeasured constructs—factors that matter for some groups but are poorly captured in existing data. Finally, they may signal greater real-world uncertainty or exposure to structural instability, particularly among marginalized groups. It is difficult to empirically disentangle these explanations using existing observational data, as doing so would likely require not just methodological adjustments but also new theory and data collection efforts tailored to group-specific sources of social stratification and unpredictability.

The measurement of social rigidity through a predictive approach is inherently shaped by the data and methods employed, just like other metrics such as the descriptive association between parental and child income or the causal estimand of the average change in child income resulting from a given change in parental income. Our analysis may be limited as we draw from a single-cohort survey, therefore our results are necessarily limited to the social rigidity in the life course progression of the late Baby Boomers. Future research could investigate whether the reported level of predictability can be generalized to other birth cohorts, especially as the NLSY97 cohort enters mid-life ages. We highlighted the need to account for two sources of uncertainty for this promising line of future research: (1) the uncertainty of the specification of model in the existing dataset, and (2) the uncertainty of adapting the model

from the earlier cohort to later cohort. For example, the younger cohort may have a higher proportion of college-educated individuals due to educational expansion. Therefore, the mechanisms for the returns on college may differ between the two cohorts. Additionally, cohorts experience labor market conditions differently; for instance, the NLSY79 cohort encountered the 2008 economic recession in their 40s, while the NLSY97 cohort faced it in their mid-20s. To address these differences, one must relax assumptions of stable relationships between predictors and outcomes across cohorts by introducing cohort interaction terms, re-estimating model parameters using partial data from the younger cohort, or adjusting for compositional differences to ensure comparability.

Our analysis is likely also limited by what variables are included in the survey. It is more ideal to capture more life experiences that are also important for mid-life SES outcomes, such as additional aspects of the college experience, institutional rankings, and neighborhood characteristics, just to name a few. However, data such as college ID and detailed addresses, which could link to data on college and neighborhood characteristics, are restricted. In contrast, we reply more on publicly available data such as high school contexts and more general measures such as regions. The impact of including these restricted-access data on predictability in this context remains an area for future research.

Lastly, fully leveraging the power of complex machine learning models often requires that the model learns from a large number of observations, especially when the feature space is big, too (Salganik et al. 2019). However, although the limited sample size may have played a role in determining predictability, we believe that an analysis of predictability using the NLSY79 data remains useful, as this type of data is a good representation of the kind of data typically available to social scientists in terms of its size and breadth (Garip 2020). It offers a valuable benchmark for future research on social rigidity and serves as a useful reference point for subsequent data collection efforts, particularly when larger sample sizes are attainable. In addition, improvements in prediction accuracy is not always guaranteed from increased sample size, as it is often jointly determined by data quality, model complexity, and computational capacity. We leave it to future studies to more systematically examine how sample sizes affect the ability to explain (in existing approaches) or to predict (in our proposed approach) socioeconomic outcomes. For example, future research could fruitfully build on this work by using other widely used data sources or larger administrative datasets (Savcisens et al. 2023). Moreover, as the machine learning methods continue to advance, employing even more innovative methods

might also help update our understanding of social rigidity through a predictive approach, such as transfer learning (Savcisens et al. 2023; Vafa et al. 2022). The core idea of such transfer learning approach is to train a predictive model—represented by a prediction function using shared covariates across datasets as inputs—on the original data and then apply it to generate predicted outcomes in the new dataset. More advanced transfer learning techniques can further adapt the prediction function to account for differences in prediction mechanisms between the original and new data. Altogether, applying predictive models to a variety of outcomes and samples can collectively improve our understanding of the intricate process of social stratification.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Haowen Zheng https://orcid.org/0000-0003-4464-9365
Siwei Cheng https://orcid.org/0000-0003-2149-0574

Data and Code Availability Statement

Data are publicly available through the National Longitudinal Surveys website maintained by the U.S. Bureau of Labor Statistics (2023). Analysis code can be accessed on the Open Science Framework at Zheng (2025) via https://osf.io/7wjg8/.

Supplemental Material

Supplemental material and Appendices for this article are available online.

Notes

1. Machine learning methods are not exclusive to the predictive approach, their use in causal inference—for estimating outcome regression functions and propensity scores—is well documented (Brand et al. 2023). However, the predictive approach we propose here more explicitly focuses on the primary objective of machine learning models: enhancing prediction accuracy.

2. These categories include, for example, "no man—mother," "father—no woman," "on my own," etc.
3. The goal is to retain as many observations as possible while reasonably preserving the information within the variables for use in machine learning models. This approach is common in non-predictive statistical analysis (Killewald and Gough 2013), where flag variables indicate variation when ambiguous values (e.g., the minimum) are assigned to missing data points. To assess robustness in the predictive setting, we also tested assigning alternative values, such as medians from the non-missing sample. As shown in Appendix A1, Supplemental Material, these adjustments do not substantively affect our main results. We opted not to drop missing data to avoid potential sample selection bias, as many variables exhibit non-random patterns of missingness.
4. The data set contains over 4,000 variables for several reasons. First, the inclusion of extensive variable lists stems from shared concepts and theories. For example, theories on school contexts and delinquent behavior each necessitate dozens of variables. We intentionally avoided using factor analysis to reduce the number of variables, as our design relies on ML models selecting from a large pool of predictors. Second, the data set spans multiple life stages, with the B set including eight waves of data. As a result, the same concepts are measured repeatedly over time. Third, to enable the ML models, we applied one-hot encoding, which transforms categorical variables with many categories (e.g., college majors and occupations) into multiple dummy variables.
5. Due to computational constraints, we used a 70/30 train-test split for the main analyses. Robustness checks in Appendix A3, Supplemental Material show using 5-fold cross-validation for selected models yield similar results. As described in the next section, we applied 5-fold cross-validation for subsequent analyses examining variable importance and cross-group predictability by gender and race.
6. We fine-tune these models using cross-validation. For the random forest model, we set the tuning parameter to the square root of the total number of predictors. We set it to this typical value, as changing the tuning parameter to other values does not lead to significantly better model performance. For boosting, we picked the number of trees by cross-validation while setting the tree depth and step size of boosting to be relatively small. For BART, we used 200 trees, 1,000 iterations, and 100 burn-in iterations. We picked these typical values as BART performs well with minimal tuning.
7. This is consistent with the finding from the current machine learning literature that compared to deep learning methods, which are well-suited for tasks such as image recognition, speech recognition, and natural language processing, tree-based models are often the preferred choice for structured tabular data sets, such as those used in our study (Chen and Guestrin 2016).

8. We used gradient boosting models for this task because SHAP works well with such tree-based methods, and that they perform similarly with random forest and BART from previous analyses. We select hyper-parameters to have a modest number of bins (512), a low learning rate (0.05), a low number of trees (10), and a modest requirement for the number of data points in a leaf node (100). We also tried other combinations (such as increasing the number of trees and decreasing the number of data points in leaves), and results are similar (see Appendix Table A3b, Supplemental Material).
9. We compared results for the main analysis before and after incorporating the 5-fold cross-validation. The results are similar (see Appendix Table A3a, Supplemental Material).
10. We choose to not standardize the number of variables for each stage as this analysis is not intended as a direct comparison across life stages (like a horse race). Additionally, simply counting the number of variables may not be meaningful, as each variable can be coded in multiple ways, leading to an arbitrary number of variables included in the model.
11. We also did a supplementary analysis using only B set variables to predict C set outcomes. Results also show that set B1 alone reaches a modest level of predictability. Those interested can refer to more details in Appendix A4, Supplemental Material.
12. Note that while traditional OLS model can accommodate interactions and polynomial terms, we have not included these non-linear terms in the linear models here. This is because OLS models do not inherently offer effective techniques for variable selection or flexible model specification that protect against overfitting, especially in comparison to machine learning methods that are better equipped to manage high-dimensional data.
13. More accurately, the need to examine income and occupational outcomes entails a sample exclusion of those who were consistently unemployed during age 40–50.
14. As aforementioned, this is not surprising as B1 includes directly many variables that could be considered as "lag variables" to outcomes in set C. For example, for outcomes that include hourly wage, total income, and family income, the "lag variables" include all individual level wage and income variables. For predicting SEI, the "lag variables" include all "SEI" variables in set B1.
15. The high social fluidity among this cohort may be partially attributed to their experiencing the 2008 economic recession during mid-life. We consider this as a significant aspect of the cohort's life experience. To further investigate, we also ran similar tasks to predict the same average outcomes between ages 36 and 44, covering pre-recession periods. The R^2_{test} increased by about 10 percentage points. This could be due to the removal of "recession effect" or that the outcomes were measured earlier in the life course before they stabilize.

16. We pick this outcome variable as an illustration because it has the highest predictability.
17. Again, the relatively large number of features is due to the transformation of some variables into dummy variables, such as the 12-category classification of parents' occupational attainments.
18. To ensure a larger sample size, which benefits our prediction task, we chose to retain older respondents in the sample, relying on age to account for these differences in the prediction model.
19. The results remain similar whether or not set B2 (adult family formation) is included. Therefore, we omit those results and instead compare outcomes before and after excluding lagged income and wage variables from set B1.
20. The maximum number of non-missing survey waves is 6, as NLSY79 switched to biennial survey schedule after 1994.

References

Athey, S. and G. Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27): 7353–7360.

Badolato, L., A. Decter-Frain, N. J. Irons, M. Miranda, E. Walk, E. Zhalieva, M. Alexander, U. Basellini, and E. Zagheni. 2023. "Predicting Individual-Level Longevity with Statistical and Machine Learning Methods."

Blau, P. M. and O. D. Duncan. 1967. The American Occupational Structure."

Bloome, D. and J. Furey. 2020. "Lifetime Inequality: Income and Occupational Differences and Dynamics in the US." *Research in Social Stratification and Mobility* 70: 100470.

Brand, J. E., R. Moore, X. Song, and Y. Xie. 2019. "Why Does Parental Divorce Lower Children's Educational Attainment? A Causal Mediation Analysis." *Sociological Science* 6: 264–292.

Brand, J. E., J. Xu, B. Koch, and P. Geraldo. 2021. "Uncovering Sociological Effect Heterogeneity Using Tree-Based Machine Learning." *Sociological Methodology* 51(2): 189–223.

Brand, J. E., X. Zhou, and Y. Xie. 2023. "Recent Developments in Causal Inference and Machine Learning."

Breen, C. F. and N. Seltzer. 2023). Sociodemographic Characteristics Alone cannot Predict Individual-Level Longevity."

Breen, C. F. and N. Seltzer. 2024. "Demographic Perspectives on Predicting Individual-Level Mortality." https://osf.io/fazsj/?view_only=cbefa9a80a684a8b9c17d1e5f8ad967. Draft version, July 31.

Browne, I. and J. Misra. 2003. "The Intersection of Gender and Race in the Labor Market." *Annual Review of Sociology* 29(1): 487–513.

Budig, M. J. and P. England. 2001. "The Wage Penalty for Motherhood." *American Sociological Review* 66(2): 204–225.

Bureau of Labor Statistics, 2023. U.S. Department of Labor. National Longitudinal Survey of Youth 1979 cohort, 1979-2020 (rounds 1-29). Produced and distributed by the Center for Human Resource Research (CHRR), The Ohio State University. Columbus, OH.

Chen, T. and C. Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." Pp. 785–794 in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*.

Cheng, S. 2014. "A Life Course Trajectory Framework for Understanding the Intracohort Pattern of Wage Inequality." *American Journal of Sociology* 120(3): 633–700.

Cheng, S. 2016. "The Accumulation of (Dis) Advantage: The Intersection of Gender and Race in the Long-Term Wage Effect of Marriage." *American Sociological Review* 81(1): 29–56.

Cheng, J., L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. 2014. "Can Cascades be Predicted?" Pp. 925–936 in *Proceedings of the 23rd International Conference on World Wide Web*.

Cheng, S., B. Chauhan, and S. Chintala. 2019. "The Rise of Programming and the Stalled Gender Revolution." *Sociological Science* 6: 321–351.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fischer, C. S., M. Hout, M. S. Jankowski, S. R. Lucas, A. Swidler, and K. Voss. 1996. *Inequality by Design: Cracking the Bell Curve Myth*. Princeton, NJ: Princeton University Press.

Frederick, C. and R. M. Hauser. 2010. "A Crosswalk for Using Pre-2000 Occupational Status and Prestige Codes with Post-2000 Occupation Codes." *Center for Demography and Ecology: University of Wisconsin-Madison. PMID*25506974. https://cde.wiscweb.wisc.edu/wpcontent/uploads/sites/839/2019/01/cde-working-paper-2010-03.pdf

Garip, F. 2020. "What Failure to Predict Life Outcomes Can Teach US." *Proceedings of the National Academy of Sciences* 117(15): 8234–8235.

Garip, F. and M. W. Macy. 2023. "Machine Learning in Sociology: Current and Future Applications."

Glauber, R. 2007. "Marriage and the Motherhood Wage Penalty Among African Americans, Hispanics, and Whites." *Journal of Marriage and Family* 69(4): 951–961.

Goldberger, A. S. 1972. "Structural Equation Methods in the Social Sciences." *Econometrica: Journal of the Econometric Society* 40(6): 979–1001.

Goldin, C. 1977. "Female Labor Force Participation: The Origin of Black and White Differences, 1870 and 1880." *The Journal of Economic History* 37(1): 87–108.

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

Imai, K. and T. Yamamoto. 2013. "Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence From Framing Experiments." *Political Analysis* 21(2): 141–171.

Killewald, A. 2013. "A Reconsideration of the Fatherhood Premium: Marriage, Coresidence, Biology, and Fathers' Wages." *American Sociological Review* 78(1): 96–116.

Killewald, A. and M. Gough. 2013. "Does Specialization Explain Marriage Penalties and Premiums?" *American Sociological Review* 78(3): 477–502.

Kim, C., C. R. Tamborini, and A. Sakamoto. 2018. "The Sources of Life Chances: Does Education, Class Category, Occupation, Or Short-Term Earnings Predict 20-Year Long-Term Earnings?" *Sociological Science* 5: 206–233.

Liu, Y. and D. B. Grusky. 2013. "The Payoff to Skill in the Third Industrial Revolution." *American Journal of Sociology* 118(5): 1330–1374.

Lundberg, S. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv preprint arXiv:1705.07874*.

Lundberg, I., R. Brown-Weinstock, S. Clampet-Lundquist, S. Pachman, T. J. Nelson, V. Yang, K. Edin, and M. J. Salganik. 2024. "The Origins of Unpredictability in Life Outcome Prediction Tasks." *Proceedings of the National Academy of Sciences* 121(24): e2322973121.

Lundberg, S. M., G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2(1): 56–67.

Mazumder, B. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *Review of Economics and Statistics* 87(2): 235–255.

McCall, L. 2005. "The Complexity of Intersectionality." *Signs: Journal of Women in Culture and Society* 30(3): 1771–1800.

Molina, M. and F. Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45: 27–45.

National Center for O*NET Development. 2022. O*NET 26.3 Database. O*NET Resource Center.

Pearl, J. 1998. "Graphs, Causality, and Structural Equation Models." *Sociological Methods & Research* 27(2): 226–284.

Pearl, J. 2012. "The Causal Mediation Formula—A Guide to the Assessment of Pathways and Mechanisms." *Prevention Science* 13: 426–436.

Reichman, N. E., J. O. Teitler, I. Garfinkel, and S. S. McLanahan. 2001. "Fragile Families: Sample and Design." *Children and Youth Services Review* 23(4-5): 303–326.

Robins, J. M. 2003. *"Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." Pp. 70–81. in Highly Structured Stochastic Systems*, edited by P. J. Green, N. L. Hjort, and S. Richardson Oxford, UK: Oxford University Press

Salganik, M. J., I. Lundberg, A. T. Kindel, C. E. Ahearn, K. Al-Ghoneim, A. Almaatouq, D. M. Altschul, J. E. Brand, N. B. Carnegie, R. J. Compton, et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117(15): 8398–8403.

Salganik, M. J., I. Lundberg, A. T. Kindel, and S. McLanahan. 2019. "Introduction to the Special Collection on the Fragile Families Challenge." *Socius* 5: 2378023119871580.

Savcisens, G., T. Eliassi-Rad, L. K. Hansen, L. H. Mortensen, L. Lilleholt, A. Rogers, I. Zettler, and S. Lehmann. 2024. "Using Sequences of Life-events to Predict Human Lives." *Nature Computational Science* 4(1): 43–56.

Sewell, W. H., A. O. Haller, and A. Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34(1): 82–92.

Stevens, G. and J. H. Cho. 1985. "Socioeconomic Indexes and the New 1980 Census Occupational Classification Scheme." *Social Science Research* 14(2): 142–168.

Sweeney, M. M. 2002. "Two Decades of Family Change: The Shifting Economic Foundations of Marriage." *American Sociological Review* 67(1): 132–147.

Vafa, K., E. Palikot, T. Du, A. Kanodia, S. Athey, and D. M. Blei. 2022. "Career: Transfer Learning for Economic Prediction of Labor Sequence Data." *arXiv preprint arXiv:2202.08370*.

VanderWeele, T. J. and E. J. T. Tchetgen. 2017. "Mediation Analysis with Time Varying Exposures and Mediators." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 79(3): 917.

Vansteelandt, S. and A. Sjolander. 2016. "Revisiting G-estimation of the Effect of a Time-Varying Exposure Subject to Time-Varying Confounding." *Epidemiologic Methods* 5(1): 37–56.

Warren, J. R., J. T. Sheridan, and R. M. Hauser. 1998. "Choosing a Measure of Occupational Standing: How Useful Are Composite Measures in Analyses of Gender Inequality in Occupational Attainment?" *Sociological Methods & Research* 27(1): 3–76.

Zheng, Haowen. 2025. Social Rigidity Across and Within Generations: A Predictive Approach. Retrieved (osf.io/7wjg8).

Zhou, X. 2022a. "Attendance, Completion, and Heterogeneous Returns to College: A Causal Mediation Approach. *Sociological Methods & Research. 53(3): 1136-1166.*

Zhou, X. 2022b. "Semiparametric estimation for causal mediation analysis with multiple causally ordered mediators." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(3): 794–821.

Author Biographies

Haowen Zheng is a PhD candidate in Sociology at Cornell University. Her research examines labor market inequality and mobility by geography, gender, and education using quantitative and computational methods.

Siwei Cheng is an associate professor of Sociology at New York University. Her research encompasses various areas of inequality, mobility, labor market, and quantitative and computational methodology.