Essays on Income Inequality

A dissertation presented

by

Oren Danieli

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Business Economics

Harvard University
Cambridge, Massachusetts

May 2019
Essays on Income Inequality

Abstract

This dissertation consists of three independent essays on income inequality. Chapter 1 (with Sydnee Caldwell) develops a method to estimate the outside employment opportunities available to each worker and to assess the impact of these outside options on wage inequality. We estimate a sufficient statistic, the “outside options index” (OOI), that captures the effect of outside options on wages, holding productivity constant. Using administrative data from Germany we find that differences in options explain 30% of the gender wage gap, 88% of the citizen-non-citizen wage gap, and 25% of the premium for higher education.

Chapter 2 (with Tanya Devi and Roland Fryer) collects new data from in-person surveys of people who grew up in poverty and develops a new approach to exploit this “wide” data in order to design the optimal experiment to study intergenerational mobility. We find that the optimal experiment should focus on educational attainment, resilience, self-esteem, locus of control, growth mindset, and trouble with the police when young.

Chapter 3 finds that the decline of middle wages in the U.S. during the 1990s (“wage polarization”) is the result of asymmetric changes in inequality at different occupations.
I show this using a new decomposition method based on the third moment (skewness), which quantifies the contributions of different factors to the total increase in wage polarization.
# Contents

Abstract ............................................................... iii

Acknowledgments ......................................................... xv

1. **Outside Options in the Labor Market** ........................................ 1
   1.1. Introduction ........................................................... 1
   1.2. A Model of Outside Options and Wages .................................... 9
      1.2.1. Setup ............................................................. 9
      1.2.2. Equilibrium ..................................................... 10
      1.2.3. Deriving an Index for Outside Options .............................. 12
      1.2.4. Sufficient Statistic ............................................... 19
      1.2.5. Parametric Example .............................................. 20
      1.2.6. Discussion ....................................................... 22
   1.3. Empirical Setting and Data ............................................... 23
      1.3.1. Data ............................................................. 23
      1.3.2. Empirical Setting: German Labor Market ............................ 25
      1.3.3. Summary Statistics ............................................... 27
   1.4. Estimating Outside Options .............................................. 29
      1.4.1. Assumptions ..................................................... 30
      1.4.2. Empirical Procedure ............................................ 32
1.4.3. OOI Input: Job and Worker Characteristics ......................... 34

1.5. The Empirical Distribution of Outside Options ....................... 38
1.5.1. Mass Layoffs ................................................. 47

1.6. Effect on Wages ................................................. 54
1.6.1. High-Speed Commuter Rail .................................. 55
1.6.2. Shift-Share (“Bartik”) ...................................... 61

1.7. Effect on Wage Inequality ......................................... 68
1.7.1. Overall Impact on the Wage Distribution ....................... 70
1.7.2. Explaining Differences in Outside Options - The Role of Commuting Costs .............................................. 72

1.8. Conclusion ....................................................... 78

2. Getting Beneath the Veil of Intergenerational Mobility: Evidence from Three Cities ........................................... 80
2.1. Introduction ....................................................... 80

2.2. A New Survey of Intergenerational Mobility in America ........... 88
2.2.1. Design ......................................................... 88
2.2.2. Sample Selection Method ................................... 89
2.2.3. Phone Screens .............................................. 89
2.2.4. Interviews .................................................... 90
2.2.5. Income and Adult Well-Being ................................ 90
2.2.6. Other Variables ............................................. 91
2.2.7. Descriptive Statistics ....................................... 96
2.2.8. Correlations with Income ................................... 98

2.3. Methods ......................................................... 102
2.3.1. An Example .................................................. 103
2.3.2. Using Observational Data to Inform Social Experiments .... 104
B.3.3. Dataset from Dobbie and Fryer 2013 .......................... 237
B.4. Technical Proofs .................................................. 238
B.5. Phone Screen Instrument ......................................... 241
B.6. In-person Interview Instrument ................................. 250
B.7. Literature Review of the Correlates of Intergenerational Mobility ........ 324
B.8. Literature Review of the Causal Impacts of Important Variables .......... 329
B.9. Appendix Figures ................................................... 336
B.10. Appendix Tables .................................................. 349

C. Appendix to Chapter 3 ................................................. 354
   C.1. Decomposing By More Categories ............................... 354
   C.2. Appendix Figures .................................................. 356
List of Tables

1.1. Descriptive Statistics ............................................................. 28
1.2. Most Weighted Question in PCA - Establishment 2014 Survey .......... 35
1.3. Most Weighted Question in PCA - BIBB ...................................... 36
1.4. Commuting Distance by Gender and Education ............................ 38
1.5. OOI by Demographics .............................................................. 46
1.6. Summary Stats for Mass-Layoff Workers by Treatment Status .......... 48
1.7. Relative Income by OOI After Mass Layoff .................................. 53
1.8. Correlation Between OOI and log wage - OLS Results ................. 55
1.9. Balance Table - High Speed Train ............................................. 59
1.10. Impact of Express Trains on Options and Wages ......................... 60
1.11. Effect of OOI on Wages Using Shift-Share (Bartik) Instrument ....... 64
1.12. Shift-Share (Bartik) Results by Exporting Share of Sales ............... 66
1.13. Shift-Share (Bartik) Results by Stayers and Movers ...................... 68
1.14. Mincer Equation with OOI ....................................................... 73
1.15. Top Standardized Values of A .................................................. 74
1.16. Distance Coefficient by Demographics ...................................... 75
3.1. Correlation of Occupational Skills .......................................... 170
List of Figures

1.1. German Districts ......................................................... 37
1.2. Distribution of Outside Option Index ............................... 39
1.3. OOI by Characteristics ............................................... 40
1.4. Cumulative Distribution of Outside Option Index by Gender .......... 42
1.5. OOI by Age ............................................................ 43
1.6. OOI Distribution by Region ........................................... 44
1.7. OOI by Training Occupation ......................................... 45
1.8. Relative Income Following Mass-Layoff .............................. 49
1.9. Mass-Layoffs - Differences in Relative Income Between High/Low OOI Workers ..................................................... 51
1.10. Mass-Layoffs - Differences in Search Time Between High/Low OOI Workers .................................................. 52
1.11. ICE Stations ............................................................. 57
1.12. Impact of Express Trains by Schooling Level ...................... 62
1.13. Shift-Share Results by Gender and Education ..................... 67
1.14. Estimates of Elasticity Between OOI and Wages ................... 69
1.15. Overall Effect on Wage Inequality ................................... 71
1.16. Effect of Commuting/Moving Costs ................................. 76
2.1. OLS Method ............................................................. 101
3.21. Share of Routine Workers by Skill ........................................ 177
B.1. Bin Scatter of Household Income with Health Indices .............. 336
B.2. Bin Scatter of Household Income with Family Based Indices ........ 337
B.3. Bin Scatter of Household Income with Childhood Experiences Indices ... 338
B.5. Bin Scatter of Household Income with Education ..................... 340
B.6. Correlations With Parental Income Controls, by Race .............. 341
B.7. Correlations With Parental Income Controls, by Gender .......... 342
B.8. Correlations With Parent Inc. Controls, by Parental Income ....... 343
B.9. Non-Parametric, by Race .................................................... 344
B.10. Non-Parametric, by Gender ............................................... 345
B.11. Non-Parametric, by Parental Income ................................... 346
B.12. Correlation Between Income and Locus of Control ............... 347
C.1. Skewness Decomposition by 3-Digit Occupation With Imputed Wages . 357
C.2. Skewness Decomposition by 3-Digit Industry .......................... 357
C.3. Skewness Decomposition by Education and Experience .......... 358
C.4. Skewness Decomposition by Occupation and Industry .......... 358
C.5. Skewness Decomposition by Occupation, School, Experience .... 359
C.6. Occupation Premium by Skill Percentile with Instruments ........ 360
I thank Chiara Farronato, Roland Fryer and Lawrence Katz for advice and guidance throughout these projects. I also thank Isaiah Andrews, David Autor, Gary Chamberlain, John Coglianese, Ashley Craig, Ellora Derenoncourt, Ed Glaeser, Nathan Hendren, Simon Jäger, Myrto Kalouptsidi, Robin Lee, Thomas Lemieux, Ro’ee Levy, Amanda Pal-lais, Assaf Romm, Jonathan Roth, Neil Thakral and Dan Svirsky for helpful comments.
1. Outside Options in the Labor Market

1.1. Introduction

In almost every model of the labor market, wages depend on a workers’ outside options: the amount of compensation they could receive from different employers. In a perfectly competitive labor market, an equally attractive outside option always exists, and competition between identical employers sets compensation at the marginal product. However, in reality, a worker’s next best option could require different skills, working hours or be located in a different city. The availability of outside options could be systematically worse for some workers due to the health of their local labor market, because they are unwilling or unable to commute, or because their skills are valuable only for a few employers or industries. Such differences could have significant implications for their incomes.

A key challenge for empirical research on this topic is that a worker’s outside option set is not typically observed. Even within the same firm and occupation, workers may face different options due to their specific set of skills, their preferences or their constraints. As a result, little is known about which workers have better outside options and what role options play in generating wage inequality.

The first contribution of this paper is to develop an empirical procedure to uncover a key latent parameter in most wage-setting models: the value of an individual’s option set. We show how this latent parameter can be derived from the cross-sectional concentration of
similar workers across jobs. If similar workers are concentrated in a certain region, industry, occupation or other job characteristics, then the worker’s options are more limited. We quantify this concentration in a single “outside options index” (OOI). We show that in a matching model of heterogeneous workers and jobs this OOI is a sufficient statistic for the effect of outside options on compensation, when holding productivity constant. We then estimate the OOI for every worker using administrative matched employer-employee data from a 1% representative sample of workers in Germany. Examining the distribution of the OOI, we find what workers’ characteristics are associated with better outside options. Next, we quantify the impact on wages by estimating the elasticity between the OOI and wages using two quasi-random sources of variation in the OOI, that holds workers’ productivity constant: the introduction of high-speed commuter-rails, and a shift-share (“Bartik”) instrument.

Our second contribution is to show that differences in outside options explain substantial portions of several widely-discussed wage gaps between different segments of workers. Outside options explain 30% of the gender wage gap in Germany. This gender difference is driven entirely by differences in willingness to commute or move. We also find that differences in outside options account for 88% of the wage gap between German citizens and non-citizens, and about 25% of the high-education wage premium. The availability of more options also increases the wage premium for urban residents. In contrast, differences in outside options reduce inequality between occupations, since high-paying occupations tend to be more specialized and workers in them therefore have fewer options.

We start by outlining a static model of the labor market that illustrates how, with two-sided heterogeneity, differences in outside options lead to differences in compensation, even for equally productive workers. Our model is based on the classic Shapley and Shubik (1971) assignment game - a two-sided matching model with transfers. Compensation
in this setting is set to prevent workers from moving to their outside options; because of heterogeneity, this will be below their full productivity in the first-best option. A direct implication is that workers’ compensation is not only determined by what they produce, but also by their ability to produce in more places.

We derive a sufficient statistic from this model, the “outside options index” (OOI), that summarizes the impact of options on compensation. It measures the quantity of relevant jobs for a given worker. If a worker gets access to more similar jobs, their compensation would increase by exactly the increase in OOI times a constant elasticity, even though their productivity remains constant. The OOI depends on two factors: the supply of jobs, and worker flexibility (i.e. a worker’s ability or willingness to take jobs in more places, more occupations, more industries, etc.). Workers with more relevant jobs, as captured by the OOI, will on average have both a better outside option, and will be able to sort into better matches, conditional on their productivity.

We show that the OOI is equal to a standard concentration index: workers with more options are those who, in equilibrium, are found in a greater variety of jobs. Under standard assumptions on the distribution of match quality (Choo and Siow, 2006; Dupuy and Galichon, 2014), the OOI is equal to the entropy index. This index, with a negative sign, is used in the industrial organization literature as a measure of market concentration (Tirole, 1988), similar to the Herfindhal-Hirschman Index (HHI), which has also been used to measure concentration in labor markets (Azar et al., 2017; Benmelech et al., 2018). In contrast to most concentration indices, our index is not measured on a specific dimension such as occupation, or industry. Instead, workers with more options are those that are less concentrated across jobs, on all dimensions included in our data set. Options here are estimated in equilibrium, based on matches we actually observe in cross-sectional data. Jobs that the worker will never take in practice because they are less attractive will not enter the OOI nor affect compensation even if the employer is willing to hire. To isolate
the effect of more options from the effect of productivity, the OOI is calculated without using any information on wages or wage offers.

We develop a method that estimates the OOI for each worker in the labor market, which is computationally feasible even in large datasets. The OOI is a function of the joint probability of every worker to be in every job. Our method estimates this probability, using the cross-sectional distribution of similar workers. We show that this problem can be translated into a logistic regression framework. We then use the fast implementation of logistic regressions to estimate the probabilities for every worker-job combination. From those probabilities we can directly calculate the OOI for each worker.

We then use the OOI to analyze the impact of outside options on inequality, starting with identification of which workers have better outside options. Specifically, we estimate the OOI for every worker in a representative sample of German workers in 2014 using administrative linked employer-employee data. Looking across observed workers’ characteristics, we find that the OOI is higher for men, German citizens, city residents, more educated and more experienced workers. We also find that higher skill workers such as medical doctors or pilots tend to be more specialized in their current industry, which narrows down their outside options. The OOI also predicts which workers will be less affected by a mass-layoff: workers with better outside options recover more quickly from a displacement. Because we do not use wages to calculate the OOI, there is not a mechanical link between the OOI and wages.

We use two sources of quasi-random variation in options, that do not affect productivity, in order to estimate the elasticity between the OOI and wages: the introduction of high-speed commuter rail stations (Heuermann and Schmieder, 2018), and a standard industry shift-share (“Bartik”) instrument (Beaudry et al., 2012). These sources of variation allow us to verify that, even if our model is not perfectly specified, there is a link between our outside options index and wages in the data. Our first source of variation focuses on
the introduction of new train stations that were constructed along existing routes. These stations effectively increased the labor market size for workers in small German cities that happened to live along the shortest route between two major cities. The second source of variation in outside options uses differences in exposure to industry growth trends between local labor markets. We compare workers who work in the same industry, but have outside options in different industries because they reside in different parts of the country. We instrument for the growth in outside options in other industries with the national industry trends to exclude the impact of local productivity shocks. Both quasi-random sources of variation yield a similar semi-elasticity of roughly .17-.32 between the OOI and wages.

Combining this elasticity with the estimated distribution of the OOI, we find that differences in outside options tend to increase wage inequality. Differences in options lower compensation for women by six percentage points, explaining roughly thirty percent of the overall gap in Germany. They also account for an eight percentage points difference in compensation between immigrants and natives, which is 88% of the overall gap. We also find large effects on the return to higher-secondary education.\(^1\) Graduates from higher-secondary education have access to more options, which increases their compensation by seven percentage points. This is about a quarter of the total return to higher-secondary education.

Finally, we examine the reasons why workers face different options. We start by examining the parameters that determine workers’ options, to understand which ones are most significant. We then use the underlying model to create counterfactual changes to the OOI under different scenarios. These exercises show that the heterogeneity in the ability to commute or move is a key factor in explaining variation in outside options. This factor can account for the full gender gap in outside options. We also find that without their

\(^1\)The level that grants a certificate allowing college admission.
higher willingness to work at more distant jobs, high-educated workers would actually have fewer options. Our analysis suggests that this is likely because their skills tend to be more industry specific.

**Related Literature** Our paper contributes to at least three distinct literatures. First, we contribute to a large literature on imperfect competition in the labor market by estimating the impact of outside options for every worker in the labor market. While outside options are a key parameter in many labor models, prior work has not focused on estimating the distribution of this parameter across different workers. Most empirical work on imperfect competition has used natural experiments in specific segments of the labor market to show that firms face upward sloping supply curves (see, e.g. Naidu, 2010; Naidu et al., 2016; Ransom and Sims, 2010; Staiger et al., 2010). Beaudry et al. (2012) and Caldwell and Harmon (2018) take a different approach, and provide direct evidence that outside options directly impact workers’ earnings, but do not investigate which workers have better options, nor the consequences for between-group wage inequality. Our paper adds to this literature by providing estimates of the distribution of options and by providing descriptive evidence on why some workers have more options than others. By combining this distribution with a causal estimate of the impact of options on wages, we are also able to present the first estimates of the impact of outside options on each individual’s wages. Our theoretical framework emphasizes market imperfections arising from worker and employer heterogeneity. This is similar to the approach taken by Card et al. (2018), and is a standard approach in the industrial organization literature for analyzing market imperfections (for instance, Berry et al., 1995). Recent work by Dube et al. (2018) shows

---

2Beaudry et al. (2012) show that shocks to one industry “spill over” onto the wages of other industries. Caldwell and Harmon (2018) show that workers with better information about their outside options see greater wage growth. Jäger et al. (2018) focus on a specific outside option—unemployment insurance—and find that changes in UI generosity has little to no effect on workers’ wages. Their result fits our finding that what matters for wage-setting is the value of a worker’s best alternative to a match. For most workers, this is likely the value of working in another job, not the value of unemployment.
that even small amounts of heterogeneity can generate substantial market imperfections. One difference between our approach and that in the search literature is that we focus on a static equilibrium. While work by Postel-Vinay and Robin (2002) shows that differences in options (as the result of on-the-job search) can impact wage growth during an employment spell, these dynamic considerations are beyond the scope of this paper.

Second, our paper contributes to a small literature on the impact of imperfect labor market competition on between-group wage inequality. Theoretical papers in this literature have argued that some groups such as women or minorities have systematically worse options, enabling their employers to pay them lower wages. These worse options may generate either higher search costs (Black, 1995) or less elastic supply to a particular firm (Robinson, 1933), and can lead to racial or gender wage gaps. Empirical papers in this literature have shown evidence that group differences do exist in both labor supply to a firm (Manning, 2003; Hirsch et al., 2010; Ransom and Oaxaca, 2010) and in rents (Card et al., 2016). A key advantage of our setting is that we are able to combine our estimates of group differences in outside options with a causally estimated elasticity between options and wages. This allows us to translate the estimated group differences in options into group differences in wages, and quantify the portion of between-group inequality that can be attributed to imperfections in the labor market.3

Finally, our paper contributes to a recent empirical literature on labor market size and concentration, by characterizing workers options using multiple worker and job characteristics at once. Manning and Petrongolo (2017) and Nimczik (2017) develop methods to uncover the size of a workers’ labor market based on willingness to commute and on observed firm-firm transitions. Azar et al. (2017) and Benmelech et al. (2018) examine trends

3Our setting expands the setting of Bidner and Sand (2016) who quantify the portion of the gender gap that can be attributed to differences in outside options driven solely by differences in access to industries. Our method includes several additional factors, such as differences in commuting costs, that we find to be generating the majority in differences in outside options between genders. We also analyze additional wage gaps beyond gender such as the education, city and citizenship premium.
in labor market concentrations by calculating Herfindahl-Hirschman indices (HHI’s) by occupation/industry, within a geographic area. Hsieh et al. (2013) estimate concentration trends by occupations and demographics such as gender and race using a model similar to ours.

In this paper we develop a method to estimate labor market size and concentration that incorporates five key features. First, when estimating workers options, we account for all job characteristics in our data. This combines all the dimensions that previous papers have used, such as geography, occupations and industry, together with job characteristics that were not used before such as working hours. Second, we account for outside options in different industries and occupations. Third, we allow each worker to have a different set of options depending on their demographics, locations, skills and preferences. Fourth, instead of assuming workers can be partitioned into distinct local labor markets, we allow options sets to overlap between workers. We also allow the distance workers are willing to travel to vary by their characteristics. Fifth, we introduce a more continuous notion of options, accounting for the fact that some options are more relevant than others. These five features allow us to estimate the value of an option set more precisely for every individual worker.

The remainder of the paper proceeds as follows: Section 1.2 outlines the theoretical matching model and derives the Outside Options Index (OOI). Section 1.3 describes the relevant features of the German labor market and the key features of the administrative linked employer-employee data that we use. Section 1.4 explains the empirical procedure of estimating the OOI. Section 1.5 describes the empirical estimates of worker outside options and presents descriptive statistics on their distribution. Section 1.6 estimates the elasticity between the outside options index and wages using two quasi-random sources of variation in options. Section 1.7 analyzes the overall effect on wage inequality. Section 1.8 concludes.
1.2. A Model of Outside Options and Wages

This section derives a model of a heterogeneous competitive labor market. We use this model to derive the outside options index (OOI), and show it is a sufficient statistic for the impact of outside options on wages. To provide additional intuition for the OOI and its effect in the model, we describe a simple parametric example. We summarize this section by discussing what is and what is not captured in the OOI using the model’s assumptions.

1.2.1. Setup

There is a continuum set of workers $I$ with measure $I$ and a continuum set of one-job firms $J$ with a measure $J$ which we pin down to 1. If a worker $i \in I$ works at job $j \in J$, they produce a value of $y_{ij}$ to the employer and a job-specific amenity valued $a_{ij}$ to the worker. The value of $y_{ij}$ is net of all costs, including capital and amenities. The value for $a_{ij}$ includes all non-pecuniary impacts on worker $i$’s utility including effort, interest, number of vacation days and more. The sum of these two values is the total value of a match, $\tau_{ij}$. This is defined for every potential worker-job pair, even those that are not observed in equilibrium. The value of $\tau_{ij}$ is taken as exogenous; all decisions by workers and employers that could affect this value such as investment in capital or human capital and location choices are pre-determined.

Employers and workers decide how to split the total surplus $\tau_{ij}$ into worker compensation ($\omega_{ij}$) and employer profits ($\pi_{ij}$).

$$\tau_{ij} = \pi_{ij} + \omega_{ij} = y_{ij} + a_{ij}$$

4Formally the value is a function $\tau : I \times J \rightarrow \mathbb{R}$.
5This is similar to Kreps and Scheinkman (1983) who show how even with competition on prices, pre-determined quantities would deviate from a Bertrand competition.
This division is accomplished via a set of transfers (wages) $w_{ij}$, which allow the worker and employer to divide the total value produced in any way between them:

$$\pi_{ij} = y_{ij} - w_{ij}$$
$$\omega_{ij} = a_{ij} + w_{ij}$$

1.2.2. Equilibrium

We next derive the allocation of workers into jobs and equilibrium wages. We use an equilibrium notion based on cooperative game theory, which is identical to the assignment game, first analyzed by Shapley and Shubik (1971). We assume a static framework with perfect information. There are additional equilibrium concepts that lead to the same result.\(^6\) We use a cooperative framework since it is somewhat more general as it does not make any assumption about how agents reach this equilibrium (e.g. who makes offers).

An allocation is defined as a set $M = \{(i, j) | i \in \mathcal{I}, j \in \mathcal{J}\}$ in which no $i$ or $j$ appears twice, so every worker can work only in one job, and every job can hire exactly one worker. For a given allocation $M$ we can define an invertible function on the domain of matched workers $m(i)$ such that $(i, m(i)) \in M$. Note that we do not require all workers and jobs to be in $M$; some workers can be unemployed and some jobs could be vacant. If a worker is unmatched, she produces $u_i$, which could be thought of as a combination of unemployment insurance and home production. Similarly, a vacant job produces $v_j$.

Shapley and Shubik (1971) show that a stable equilibrium (core allocation) includes an

\(^6\)Pérez-Castrillo and Sotomayor (2002) show one specific mechanism that leads to the same equilibrium using sub-game perfect Nash equilibrium.
allocation $M$, and a transfer $w_{ij}$ for each $(i, j) \in M$ which satisfies

$$\forall i', j' \in I, J : \omega_{i'} + \pi_{j'} \geq \tau_{i' j'}$$

(1.1)

$$\forall i' \in I : \omega_{i'} \geq u_{i'}$$

$$\forall j' \in J : \pi_{j'} \geq v_{j'}$$

where $\omega_{i'} = \omega_{i', m(i')}$ if worker is matched and $\omega_{i'} = u_{i'}$ otherwise, and similarly $\pi_{j'} = \pi_{m-1(j') j'}$ or $v_{j'}$.

The first condition says that there is no single worker-employer combination that could deviate from their current allocation, produce together, and split the surplus in such a way that both the employer and the worker would be better off. Note that this condition includes all possible combinations, including those that are not matched in equilibrium. The second and third conditions are participation constraints which require that every worker and employer obtain no less than their unemployment or vacancy value.\footnote{Unemployment and vacancies can exist simultaneously, as long as $u_i + v_j \geq \tau_{ij}$ for every possible match of non-participants.} Shapley and Shubik (1971) shows that a stable allocation $M^*$ is also optimal in the sense that the maximum total value is produced.

Workers’ compensation in this model depends not only on the value they produce in their workplace, but also on the value they produce in other jobs. Compensation is strictly bounded by the worker and the employer’s marginal contributions to the entire market (Roth and Sotomayor, 1992). Because this marginal contribution to the market is weakly smaller than the productivity at the workplace, workers are paid below their full productivity. Workers who are able to produce a similar value in more places will get a larger portion of their productivity to keep the equilibrium stable.
1.2.3. Deriving an Index for Outside Options

We next examine the role of outside options in this equilibrium. In particular, we derive the outside options index (OOI), a sufficient statistic for the impact of outside options on wages.

In any stable equilibrium, each worker must earn more in her current match than she could earn at a different employer.

\[ \omega_{ij} \geq \max_{j' \neq j} \omega_{ij'} \]  

(1.2)

This outside option \( \omega_{ij'} \) is exactly what will make employer \( j' \) indifferent between their equilibrium match, and hiring \( i \) (formally, \( \tau_{ij'} - \omega_{ij'} \geq \pi_{j'} \)). Hence

\[ \omega_{ij'} = \underbrace{\tau_{ij'}}_{\text{potential \ j' equilibrium}} - \underbrace{\pi_{j'}}_{\text{value \ i, j' compensation}} \]  

(1.3)

Combined we get a lower-bound for worker compensation\(^8\)

\[ \omega_{ij} \geq \max_{j' \neq j} \tau_{ij'} - \pi_{j'} \]  

(1.4)

The employer decision can thus be written as the solution to a simple profit maximization problem.\(^9\)

From these equations, we can derive an expression for worker’s compensation that we

---

\(^8\)Equilibrium compensation \( \omega_i \) must satisfy \( \omega_i + \pi_{j'} \geq \tau_{ij'} \), yielding this equation. This bound will be tight as long as \( \max_{j' \neq j} \tau_{ij'} - \pi_{j'} \geq u_i \). It holds with equality under an additional assumption (Assumption 1.1).

\(^9\)Equation 1.2 defines the effective price that employer \( j \) needs to pay to hire worker \( i \). In order to maximize profit, the employer needs to choose a worker that maximizes the value net of cost: \( \max_{j'} \pi_{ij'} = \max_{j'} \tau_{ij'} - \omega_{ij'} \).
can take to the data. First, define $X \subseteq \mathbb{R}^{d_x}, Z \subseteq \mathbb{R}^{d_z}$ to be the characteristic spaces of workers and jobs accordingly. Let $X_i$ and $Z_j$ denote the observed worker and job characteristics which are distributed with a density $f (X_i), f (Z_j)$ respectively.\(^{10}\) We next add an assumption on the distribution of $\tau_{ij}$ based on these observables. We follow Dupuy and Galichon (2014) and assume that the value of $\tau_{ij}$ conditional on the observables is drawn from a sum of two continuous logit models, one for the workers and one for the employers. This is a generalization of the classic multinomial logit for a continuous case.

**Assumption 1.1.** The match value $\tau_{ij}$ between a worker with observable characteristics $x_i$, and a job with observable characteristics $z_j$, can be written as

$$
\tau_{ij} = \tau (x_i, z_j) + \varepsilon_{ij}
$$

where $\varepsilon_{ij}$ has the following distribution

$$
\varepsilon_{ij} = \varepsilon_{i,z_j} + \varepsilon_{j,x_i}
$$

s.t. $\varepsilon_{i,z_j} \perp \varepsilon_{j,x_i}$

$$
\varepsilon_{i,z_j}, \varepsilon_{j,x_i} \sim CL (\alpha)
$$

$CL (\alpha)$ is the continuous logit distribution, that closely resembles an extremum value type-1 distribution with scale $\alpha$ (Dagsvik, 1994). For details on this distribution, see Appendix A.1.1.\(^{11}\)

This assumption simplifies the math considerably. However, it is strong; it implies that workers have an unobserved utility or productivity in jobs with specific observed characteristics, and those unobserved shocks are uncorrelated, even between jobs with similar

---

\(^{10}\)Formally, there are measurable functions $X_i : I \to X, Z_j : J \to Z$.

\(^{11}\)Formally, every worker $i$ draws $\varepsilon_{i,z_i}$ shocks from a Poisson process on $Z \times \mathbb{R}$. As a result, for every subset $S \subseteq Z$, $\max_{z \in S} \{ \varepsilon_{i,z} \} \sim EV_1 (\alpha \log P (S) + \text{const}, \alpha)$ where $EV_1$ is extremum-value type-1 distribution, and $P (S) = \int_S f (z) \, dz$. A similar process exists for $\varepsilon_{j,x_i}$. More details in Appendix A.1.1.
characteristics. Employers also have similar unobserved independent shocks based on the workers observables. Moreover, the assumption that $\varepsilon_{i,z,j} \perp \varepsilon_{j,x,i}$ implies that there are no interactions between the worker and job unobserved characteristics.

We can rewrite the latent value of outside options from Equation 1.3 as

$$\omega_{ij'} = \tau(x_i, z_{j'}) - \pi_{j'} + \varepsilon_{ij'}$$

(1.5)

Using $^*$ to denote the best alternative offer ($\omega_{ij'}^* = \tau^*(x_i, z_{j'}) - \pi_{j'}^* + \varepsilon_{ij'}^* = \max_{j'} \omega_{ij'}$), we get a simple expression for the expected value of the best alternative offer:

$$E[\omega_{ij'}^*] = E[\tau^*(x_i, z_{j'})] - E[\pi_{j'}] + E[\varepsilon_{ij'}^*]$$

(1.6)

This decomposition is the key result of our theoretical analysis.

The first component reflects the mean value the worker can produce where they typically work (without strategic sorting on $\varepsilon^{12}$). Therefore, as in almost all labor models, workers that produce a higher value would earn a larger compensation. The second component reflects the mean employer profit, beyond costs. This could be zero, or constant if we think the employer market clears perfectly through entry, but we do not assume this is necessarily the case. This component is affected by several factors including the firm productivity, and the market price of their workers.

In this paper, we focus on the third component $E[\varepsilon_{ij'}^*]$. This expression depends on the measure of relevant options the worker has. Even though for a random match, $E[\varepsilon_{ij}]$ is constant, the expectation of $\varepsilon_{ij'}^*$ is higher because $j'$ is positively selected, since it is the second best option. The more similar options in expectation worker $i$ will have, the larger this component will be.

---

12 $\tau(x_i, z_j)$ reflects the expected value a worker can produce in a random job with characteristics $z_j$. 

14
We derive the Outside Option Index, OOI, directly from this expectation. Specifically, we define the OOI to be the standardized expectation $\frac{1}{2\alpha} E \left[ \varepsilon_{ij}^* \right]$, where $\alpha \geq 0$ is the scale parameter that depends on the distribution of $\varepsilon_{ij}$. This descaling guarantees that the value of OOI is independent of the standard deviation of $\varepsilon$, and therefore of the units in which we define $\tau(x, z)$. $\alpha$ also sets the link between the OOI and wages, which we’ll estimate using two distinct quasi-random sources of variation in options in Section 1.6. We assume that $\alpha$ is constant across all workers, implying a constant elasticity between OOI and wages. Our results on the heterogeneous impact of options on wages in Section 1.6 are consistent with this assumption.

The standard result that workers earn what they produce is a particular case of this setting. This occurs when $\alpha = 0$ ($\varepsilon_{ij}$ is constant at 0) and entry decision of employers are optimal, such that profit is zero. This emphasizes the key distinction of this more general setting from the perfectly competitive model: heterogeneity. When $\alpha > 0$, there is no identical employer to bid wages up to the worker’s full product.

Under Assumption 1.1, workers and employers are indifferent between matches with the same characteristics. Formally, defining $f^i_j$ the probability density of worker $i$ to work at job $j$ in equilibrium we get the following lemma:

**Lemma 1.1.** Under Assumption 1.1, the probability density of worker $i$ to work at job $j$ satisfies

$$
    f^i_j = \frac{f(X_i, Z_j)}{f(X_i)} \frac{f(Z_j)}{f(X_i)}
$$

where $f(X_i, Z_j)$ is the joint density of matched worker and job observables in equilibrium.

Intuitively, this lemma implies that $f^i_j$ is equal for all jobs with the same characteristics. Appendix A.1.2 provides a full proof, as well as formal definitions for those densities.

This assumption also yields a closed-form expression for the outside options defined in
Equation 1.5.

Lemma 1.2. Under Assumptions 1.1, in equilibrium, worker $i$ with characteristics $x_i$ is facing a continuous logit choice between employers who are offering

$$
\max_{z_j} \omega (x_i, z_j) + \varepsilon_{i,z_j}
$$

and

$$
\omega (x_i, z_j) = \tau (x_i, z_j) - \pi (z_j) - \alpha \log f^i_j
$$

where $\pi (z_j) = E [\pi_{j'} | Z_j = z_j]$. Similarly, employers choose between

$$
\max_{x_i} \pi (x_i, z_j) + \varepsilon_{j,x_i}
$$

This lemma simplifies the matching procedure into two one-sided continuous logit choices. Because employers with the same characteristic $z_j$ are willing to make the same offer, the best alternative offer $\omega^*$ equals the maximal offer the equilibrium employer is willing to make. Hence, the lower bound from Equation 1.4 can be replaced with an equality. The workers are facing a choice between the employers who are looking for workers with their observed characteristics $x_i$.

The market clears when the supply of workers with characteristics $x_0$ to jobs with characteristics $z_0$ equals demand. Demand is decreasing with quantity, because the marginal employer has a lower value of $\varepsilon_{j,x_0}$. This is why the compensation $\omega_{ij'}$ in lemma 1.2 depends negatively on $\alpha \log f^i_j$. However, supply increases with compensation, hence $f^i_j$ will be increasing in $\tau (x_i, z_j) - \pi (z_j) - \alpha \log f^i_j$. Those two equalize exactly when

$$
f^i_j \propto \exp \frac{1}{2\alpha} [\tau (x_i, z_j) - \pi (z_j)]
$$
This result implies that we can learn about the quality of outside options based on how similar workers sort into different jobs. Workers tend to sort into jobs where their net productivity is highest, where net productivity is the difference between the mean value they can produce $\tau(x_i, z_j)$, minus the employer’s expected profits $\pi(z_j)$. Therefore, the only valuable outside options for a worker are those that are taken, in equilibrium, by similar workers.

With this distributional assumption we can find an analytical expression for the OOI. The expected value of $\varepsilon_{ij}^*$ simplifies with the following lemma

**Lemma 1.3.** Under Assumption 1.1:

$$E[\varepsilon_{ij}^*] = E[\varepsilon_{i,zj}^* + \varepsilon_{j,xi}^*] = -2\alpha \int f_j^i \log f_j^i \quad (1.7)$$

The last equality follows because both $\varepsilon_{i,z}, \varepsilon_{j,x}$ are drawn from a continuous logit with scale parameter $\alpha$. To measure the OOI for worker $i$, we need to take the integral over all their potential matches. Using the definition in Equation 1.6 yields

$$OOI = \frac{1}{2\alpha} E[\varepsilon_{ij}^*] = -\int f_j^i \log f_j^i \quad (1.8)$$

This expression is the well-known entropy index. Entropy is frequently used to measure industry concentration. Analogously, the OOI can be thought as a concentration index across jobs. A worker with more options (a worker who is less concentrated) will have a higher OOI because their probability of being in a specific job is lower. This is concentration on all (observable) dimensions: location, occupation, industry etc. Empirically, we will estimate it based on the concentration of workers with similar observables. If similar workers tend to be concentrated in a specific region of the country, small number of occupations or industries we will estimate a lower OOI for them. We will describe this
procedure in detail in Section 1.4.

The entropy index is also commonly used in measuring unpredictability. In our context, this would be the difficulty to predict the worker’s job. Workers whose jobs are harder to predict, are those with more options. The OOI takes values on \((-\infty, 0]\). As the measure of jobs a worker can take approaches zero, \(OOI \rightarrow -\infty\); if a worker is equally likely to take any job, \(OOI = 0\).

This OOI is driven by two factors. First is worker flexibility, the ability of the worker to take jobs at different locations, use their skills in different occupations, industries etc. All of which we will measure empirically. Second is the supply of relevant jobs. More relevant jobs that the worker can take will increase the OOI directly. The OOI is only driven by relevant outside options – jobs that similar workers are actually observed taking in equilibrium \((f_j^i > 0)\). Empirically, this will be options that are actually sometimes executed by workers with similar observables. Therefore, jobs that a worker could do but never would do in practice will not enter the OOI, and won’t affect the equilibrium outcome.

The key advantage of using the OOI is that it does not depend on any information on a worker’s alternative wages. This is useful because information on potential wages at other jobs is typically unavailable. Moreover, a worker’s alternative wages also depend directly on their productivity. The OOI captures the impact of options on wages, holding productivity constant. This also implies that any link that we find between the OOI and wages is not mechanical.

\footnote{It is possible that it’s easier to predict the job of certain workers due to better data quality. This would imply that those workers will have a lower scaling parameter \(\alpha\), and therefore a lower elasticity between the OOI and wages. To test this, we estimate the OOI-wage elasticity in Section 1.6 separately by gender and education. Our results are consistent with a constant value of \(\alpha\) for all workers.}
1.2.4. Sufficient Statistic

The OOI is a sufficient statistic for the effect of access to more options on workers compensation, under our model assumptions. Access to options has two distinct effects on workers, both of which are captured in the OOI. First, it improves workers compensation at the same job by improving their outside options. Second, the improvement in options allows some workers to find better matches.

We first define an improvement in access to options. We define $\lambda_x$ to be the measure of a random set of jobs that are accessible to workers with observables $x$. All jobs that are not accessible have $\tau_{ij} = -\infty$ and are therefore never chosen in equilibrium. We model an increase in access to more jobs would be an increase to this $\lambda_x$. In Appendix A.1.2 we show that other definition of $\lambda$ such as a linear commuting cost would yield the same results.

Theorem 1.1 shows that workers who get access to more outside options get an increased wage offer from their employer that equals to $\alpha$ times the change in their OOI.

**Theorem 1.1.** Let $j$ be $i$’s equilibrium match. Access to outside options $\lambda_x$ has the following effect on the maximum offer $j$ is willing to make in the new equilibrium:

$$\frac{d\omega_{i,j}}{d\lambda_x} = \alpha \frac{d\text{OOI}}{d\lambda_x}$$

The second effect of access to more options is an improvement in match quality. An improvement in outside options is only an improvement, if some workers would in practice match into those additional jobs in equilibrium. Therefore, the overall effect of access to more options is a combination of the better outside options, and the option to improve match quality. The following theorem shows that in this model, the overall effect is exactly twice the size of the effect only through outside options.
**Theorem 1.2.** Access to options $\lambda_{x_i}$ has the following overall effect on expected worker compensation in equilibrium

$$\frac{dE[\omega_{i,j}]}{d\lambda_{x_i}} = 2\alpha \frac{dOOI}{d\lambda_{x_i}}$$

Different choices of counterfactuals could potentially lead to different results. The counterfactual we consider is giving a small group of workers access to more similar jobs. If the increase in $\lambda$ affects a non-zero measure of workers, then there will be general equilibrium impacts on employer profits. For instance, mandating stable working hours in all jobs will give all women access to more jobs. This counterfactual may decrease the profits of employers that were already hiring mostly women. In such cases, the OOI would only be a sufficient statistic if the employers’ market is perfectly competitive such that profit are kept constant through entry and exit. Access to better jobs (as opposed to similar jobs), in which the worker can produce greater value will also affect workers productivity, and therefore will affect compensation beyond the effect on the OOI.

### 1.2.5. Parametric Example

To give further intuition for the OOI, and the additional components in our key decomposition (Equation 1.6) we go over a simple parametric example.

In this simple setting, workers are characterized only by their productivity and their amount of options. Assume workers and jobs are equally dispersed across the real line $\mathbb{R}$. Each worker can be described as a 3-dimensional tuple $(l_i, y_i, d_i)$ which is her location on the real line, her productivity and the maximal distance she is able to commute. Jobs

---

14Formally, assume each interval $[a, b]$ has a measure of $b - a$ workers and jobs. This implies an infinite measure of both workers and jobs.
are identical other than their location $l_j$. The value of a match is then

$$\tau_{ij} = \begin{cases} y_i + \varepsilon_{ij} & |l_i - l_j| < d_i \\ -\infty & \text{else} \end{cases}$$

where $\varepsilon_{ij}$ are the sum of two continuous logit distribution as before.

In this simple setting, the OOI corresponds to the log measure of options. The PDF of a worker distribution across jobs is constant at $\frac{1}{2d_i}$ for all jobs within feasible range. Therefore, the OOI is $-\log \frac{1}{2d_i} = \log 2d_i$ which is the log of the measure of jobs a worker can take. Differences in OOI are therefore the log ratio in the measure of relevant options. This result will generally hold for every pair of workers with similar distribution of jobs and different sizes of support, not only in this example. In this setting, $\lambda$ is exactly $2d$, hence from Theorems 1.1 and 1.2, an infinitesimal increase in $d_i$ leads to an increase of $\frac{\alpha}{2d_i}$ if they stay at the same job, and $\frac{\alpha}{d_i}$ overall.

The first component of Equation 1.6 (mean value) captures a worker’s baseline productivity; in this case this is equal to $y_i$. This component represents the expected productivity in a random job that a worker could take. Equivalently, it captures productivity differences, conditional on having the same amount of options (OOI). The final component, employer rents, will be equal for all workers, as all jobs are equivalent.\(^{15}\)

This example shows clearly how two workers who are on average equally productive, could still earn different wages due to differences in outside options. Assume $l_1 = l_2$, $y_1 = y_2$, and $d_1 < d_2$. Worker 2 earns a higher wage because her OOI is greater. In expectation, workers 1 and 2 are equally productive at every job in $[l - d_1, l + d_1]$. Since worker 2 has a higher price, most jobs in this range would prefer to hire worker 1. Still, as a result of heterogeneity, some employers would be willing to pay the higher price.

\(^{15}\)Its exact value would be pinned down depending on the value of unemployment, and vacant jobs.
Because worker 2 has more options than worker 1, there are enough employers who are willing to pay the higher price, so that the market clears.

1.2.6. Discussion

We summarize this section by re-examining the model assumptions and their implication on what is and what is not captured with the OOI. The primary advantage of the OOI is that it more precisely captures the size of a worker’s relevant option set. It allows workers to use their skills in different occupations and industries. By contrast, measures such as the HHI assume that workers belong to only one industry or occupation. Similarly, the OOI accounts for heterogeneity in commuting and moving costs. Instead of assuming each worker is assigned to a specific local labor market, the OOI empirically assess the distance over which each type of worker searches for a job. Finally, the OOI accounts for variation in employer characteristics even within the same industry. For instance, if some workers are unable to work on weekends, their OOI will only be affected by employers who do not require that.

The main limitation of the OOI is that it does not account for any dynamic considerations. This is because it was derived from a static model. Dynamic considerations such as switching costs, firm-specific human capital that is acquired over time, and learning tend to limit a worker’s ability to move to their outside options, but are beyond the scope of this analysis.

A second limitation is that the OOI calculates the measure of relevant jobs, not relevant employers. We assumed that employers are 1-job firms and do not account for the fact that many jobs are under the same employer. While the model will, with minor adjustments, accommodate firms, we focused on jobs due to limitations of our data (see Section 1.3.1). Therefore, the OOI will over-estimate options for workers who are more likely to work in
large firms.

In contrast, some aspects of the labor market that are not explicitly modeled above could still be captured in the OOI. The most prominent one is information frictions that would generate search costs. Black (1995) has analyzed a search model where some workers have more options, and showed that in this setting as well, more options would lead to higher wages in equilibrium. Hence, it is possible that some of the effect of the OOI on wages is operating through this channel as well.

1.3. Empirical Setting and Data

We use administrative data from Germany to generate measures of individual workers’ outside options. The data includes detailed information on establishment and worker characteristics, including information on a variety of amenities provided by different establishments, which allow us to estimate workers’ options more accurately. Excluding some idiosyncratic features which we will now discuss, the German labor market is comparable to other low-regulated labor markets, making wages more directly affected by the market forces we want to study.

1.3.1. Data

1.3.1.1. Administrative German Employer-Employee Data

Our primary source of data is a panel of German worker employment histories known as the “LIAB Longitudinal” dataset. It is a matched employer-employee administrative data, based on a sample from the universe of German Social Security records from 1993-2014. There are four key features of the data which make it ideal for our setting. First, it
is a large dataset, including about 1% representative sample of the entire German labor force. Second, there is detailed establishment-level survey information with information such as hours requirements, profitability, leave/maternity policies etc. This allows us to account for differences in outside options that may be due to differences between establishments, even within industries. Third, the panel structure of the data, allows us to track workers over long periods of time. This gives us valuable information about the workers such as their specific experience in the market, and their location before taking their job. Fourth, this data provides 4-digit occupational classification which highly improves our precision in measuring relevant options. To our knowledge, this combination of data is not available in the United States.

The data come from the Integrated Employment Biographies (IEB) dataset, which is collected by the German Institute for Employment Research (IAB). Employers are required to report daily earnings (subject to a censoring limit at the maximum taxable earning level)\(^\text{16}\), education, occupation, and demographics for each of their employees at least once per year, and at the beginning of any new employment spell. New spells can arise due to changes in job status (e.g. part-time to full-time), establishment, or occupation.

Each year the IAB selects a stratified random sample of establishments from the pool of all German establishments with at least one employee liable to Social Security. These establishments are required to complete a series of surveys on organizational structure, personnel policies, financing, and research activities. In particular, the establishments are asked for information on their annual sales, profits, establishment size and leave policies. The survey data are then merged with the complete employment histories of all individuals who worked at least one day in any of these firms between 1993 and 2014.

There are several limitations for the data, that may affect our calculations of outside options. The data do not cover civil servants or the self-employed, which comprise 18% of

\(^\text{16}\)11% of the sample is censored. As we do not use wages to calculate OOI, it is not affected by censoring.
the German workforce. They also do not cover labor force non-participants. Therefore, we do not account for any of those options when calculating the OOI. Since the sample is done at the establishment level, we usually observe only few establishments in each industry-region combination. This is why we construct the OOI at the job level, and not the employer level.

Because our model is static, we rely on repeated cross-sections of data. For each year, we use data on employment relations on June 30th of each year. Our descriptive analysis is done for our last year in the sample, that is June 30th 2014. We use data from 1999, 2004 and 2012 in Section 1.6 to examine how quasi-random variations in the OOI effect wages.

1.3.1.2. BIBB Task Data

We supplement these data with survey information on the characteristics of occupations and industries. It includes information on the tasks completed, hours requirements and typical working conditions in these occupations/industries. These data are similar to the O*NET series, but allow us to account for possible differences in the task content of occupations between the United States and Germany, as well as differences in coding.\(^\text{17}\) The survey is conducted by the IAB and includes information on respondents’ occupation, industry, in addition to responses on questions related to organizational information, job tasks, job skill requirements, health and working conditions.

1.3.2. Empirical Setting: German Labor Market

There are several distinctive features of the German labor market which are relevant for our analysis. First, there are different levels of secondary-school leaving certificates,
which depend on the number of years and type of education. Our data allows us to distinguish between three categories: lower-secondary, which typically requires nine years of schooling, intermediate-secondary, which typically requires ten years of schooling, and high-secondary, which requires twelve to thirteen years of schooling, and allows the student to pursue a university degree. In our analysis we use indicators for the type of secondary education to account for years of schooling, and school quality.

Second, in addition to (or sometimes instead of) formal education, many German workers receive on-the-job training through formal apprenticeships. Individuals in apprenticeship programs complete a prescribed curriculum and obtain occupation-specific certifications (e.g. piano maker). We use this information to precisely identify the types of jobs a worker could perform.

Third, eleven percent of workers in Germany work under “fixed-term contracts” (as of 2014). These contracts expire automatically without dismissal at the end of the agreed term, at no cost to the employer. The maximal period for employment under such contract varies between 6 to 18 months over the period for which we have data. At the end of a contract, the worker and employer may choose to continue the employment relationship, but cannot use another fixed term contract to do so (Hagen, 2003).

Fourth, two percent of workers are hired through temporary work agencies. This is a triangular employment relationship, which involves the temporary work agency, a client company and a temporary worker. Historically these working relations were limited to 24 months; their duration is no longer regulated. There are additional regulations on the pay received by workers hired through temporary agencies (in particular relating to how these workers are paid relative to other workers at the same firms) but the rules vary significantly over time. In our analysis, we distinguish between employment found via temporary work agency and work found via more traditional means (Mitlacher, 2008).
While wage setting in Germany was historically governed by strong collective bargaining agreements, employers today have considerable latitude in setting pay (Dustmann et al., 2009). While employers could always raise wages above the agreed-upon levels, it only became common for contracts to include “opening clauses” allowing employers to negotiate directly with workers to pay below-CBA wages in the 1990s. Today these clauses are very common.

1.3.3. Summary Statistics

Table 1.1 describe the characteristics of workers and jobs in our sample, for the full sample, as well as by gender.

Our sample is roughly evenly split between male and female workers. The mean age for a worker in our sample is forty-five years old and the vast majority (97%) are citizens. The workers are divided about equally between the three types of secondary education. In nineteen percent of the sample the lower- and intermediate- secondary education categories are aggregated.\(^{18}\) Men and women have similar age, education and citizenship status.

On the job side, thirty one percent of the jobs in our sample are part-time. Eleven percent of jobs are on fixed contracts and only two percent are from temporary agencies. The distribution of establishment size is very skewed, with mean of 1,552 workers and a standard deviation of five times that size. The mean annual sales per worker are 163,000 Euros. Twenty-six percent of the establishments report to have females in managerial positions.

It can already be observed that men and women sort into different types of jobs. Females are much more likely to work in part-time jobs (53% compared to 13%), which is relatively

\(^{18}\)More details in data Appendix A.2.1
Table 1.1.: Descriptive Statistics

(a) Workers

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Female</td>
<td>.46 (.50)</td>
<td>1 (.00)</td>
<td>0 (.00)</td>
</tr>
<tr>
<td>Age</td>
<td>45.05 (12.49)</td>
<td>45.53 (12.27)</td>
<td>44.66 (12.67)</td>
</tr>
<tr>
<td>German Citizen</td>
<td>.97 (.17)</td>
<td>.98 (.15)</td>
<td>.96 (.19)</td>
</tr>
<tr>
<td>Education: Higher Secondary</td>
<td>.29 (.45)</td>
<td>.30 (.46)</td>
<td>.28 (.45)</td>
</tr>
<tr>
<td>Education: Intermediate Secondary</td>
<td>.31 (.46)</td>
<td>.34 (.47)</td>
<td>.28 (.45)</td>
</tr>
<tr>
<td>Education: Lower Secondary</td>
<td>.22 (.41)</td>
<td>.16 (.37)</td>
<td>.26 (.44)</td>
</tr>
<tr>
<td>Education: Intermediate/Lower</td>
<td>.19 (.39)</td>
<td>.20 (.40)</td>
<td>.18 (.38)</td>
</tr>
<tr>
<td>N</td>
<td>450,917</td>
<td>162,780</td>
<td>288,137</td>
</tr>
</tbody>
</table>

(b) Jobs

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Part - Time</td>
<td>.31 (.46)</td>
<td>.53 (.50)</td>
<td>.13 (.34)</td>
</tr>
<tr>
<td>Fixed Contract</td>
<td>.11 (.31)</td>
<td>.11 (.32)</td>
<td>.10 (.30)</td>
</tr>
<tr>
<td>Temporary Agency</td>
<td>.02 (.12)</td>
<td>.01 (.08)</td>
<td>.02 (.15)</td>
</tr>
<tr>
<td>Establishment Size</td>
<td>1,552.8 (7,679)</td>
<td>827.2 (5,014)</td>
<td>2,166.2 (9,313)</td>
</tr>
<tr>
<td>Annual Sales per worker (Euro)</td>
<td>163,286 (185,651)</td>
<td>130,414 (163,955)</td>
<td>191,026 (197,953)</td>
</tr>
<tr>
<td>%Female in Management</td>
<td>.26 (.31)</td>
<td>.36 (.35)</td>
<td>.17 (.24)</td>
</tr>
<tr>
<td>N</td>
<td>450,917</td>
<td>162,780</td>
<td>288,137</td>
</tr>
<tr>
<td>N Establishments</td>
<td>8,792</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: This table shows summary statistics of all workers and jobs in our sample on June 30th 2014. Sampling weights are used to make this a representative sample of the German population.
high compared to other countries.\textsuperscript{19} They also work at smaller establishments with 827 employees on average and mean annual sales of 130,000 Euros, compared to 2,166 and 191,000 accordingly for males. Females are also more concentrated at establishments with higher share of female-management (36\% compared to 17\%).

1.4. Estimating Outside Options

In this section we describe how we estimate the outside options index. Our method uses the cross-sectional allocation of observably similar workers to estimate the relevant options of each worker. This allocation teaches us about the worker’s ability or willingness to commute, about the set of industries or occupations that are suitable to the worker’s skills, and about the worker’s demand for certain workplace amenities. Section 1.4.1 states the key assumption, Section 1.4.2 describes the estimation procedure, and Section 1.4.3 describes the worker and job characteristic we use as inputs.

The OOI of a given worker requires an estimate of their probability to work in each one of the jobs observed in the data. We calculate the OOI using Equation 1.8 which shows that the OOI is only a function of the different $f_j^i$. This requires us to estimate $N^2$ distinct probabilities.

Earlier methods that were developed to estimate such densities do not work on data sets of our size. Non-parametric approaches cannot be used due to the large number of worker and firm characteristics. Parametric methods that were designed specifically for this model, work well when the number of possible combinations is around a few millions.\textsuperscript{20} However, given the size of our data, we need to calculate the probability density

\textsuperscript{19}Germany is ranked 6th out of 35 OECD countries in female part-time employment (OECD, 2018b).

\textsuperscript{20}Choo and Siow (2006) develop a non-parametric method where the number of possible combinations is finite and small. Dupuy and Galichon (2014) use Iterative-Proportional-Fitting algorithm to estimate a continuous density. Their data set had about $N^2 = 10^6$ possible combinations.
of approximately 250 billion possible combinations, making these methods computationally not feasible.

To overcome this challenge, we develop a new method that is computationally feasible for large data sets, and uses a similar set of assumptions to those used in the prior literature (Dupuy and Galichon, 2014). Our method relies on an equivalent representation of the probability densities as the ratio between the likelihood of a matched pair to appear in the equilibrium allocation compared to a random one. These ratios can be estimated quickly using logistic regressions. We discuss the links and differences between our method and prior methods in more detail in Appendix A.1.3.

1.4.1. Assumptions

In this section we state the parametric assumptions we make to link the $f_{ij}$ densities to the data. Our data are comprised of pairs of matches between workers, and jobs $(x_k, z_k)$, where the $x_k / z_k$ are observed worker/job characteristics we discuss in the Section 1.4.3. We first use the result of Lemma 1.1

$$ f_{ij} = \frac{f(X_i, Z_j)}{f(X_i) f(Z_j)} $$

$f(X_i, Z_j)$ is the probability of observing a match between a worker with characteristics $X_i$ and a job with characteristics $Z_j$. $f(X_i) f(Z_j)$ is the product of two the marginal distributions for workers and job characteristics. This is the probability of observing a match with such observables, under a random assignment. The basic intuition for this result is that the probability of observing $i$ matched with $j$ depends on the frequency that workers and jobs with such observables are matched, accounting for the total measure of workers and jobs with these observables (if there are more jobs with a particular set of observables, the
probability to match to a specific one is smaller). This result can be derived from weaker assumptions as well.\textsuperscript{21}

Our second assumption parametrizes $f_j^i$ as a function of the observables. We follow Dupuy and Galichon (2014) in assuming that the log density is linear in the interaction of worker and job characteristics.

**Assumption 1.2.** The log of the probability density is linear in the interaction of every worker and job characteristic:

$$\log f_j^i = X_i A Z_j + a(X_i) + b(Z_j)$$

The matrix $A$ includes all the coefficients on each of the interactions between worker and job characteristics. The marginal distributions $f(x), f(z)$ are fully determined by $a(x)$ and $b(z)$.\textsuperscript{22}

This assumption reduces the dimension of the problem significantly, while allowing the relationship between each pair of covariates to remain unrestricted. Dupuy and Galichon (2014) show that $A$ is proportional to the cross-derivative of $\tau$

$$2\alpha A = \frac{\partial^2 \tau}{\partial x \partial z} \quad (1.9)$$

where $\alpha$ is the scale parameter of $\varepsilon$ we defined in Assumption 1.1. Intuitively, this means that if a worker characteristic and a job characteristic are complements, they will be observed more frequently in the data.

\textsuperscript{21}It is sufficient to assume If $X_i = X_i'$ and $Z_j = Z_j'$ then $f_j^i = f_j'^i$, instead of Assumption 1.1.

\textsuperscript{22}While Dupuy Galichon are able to fit the marginal distribution precisely to their observed value in the data, we won’t be able to this with our data size. Therefore, we take linear functions of all $X$ variables and $Z$ variables. We also include indicators for district. As we discuss in Appendix A.1.3 this specification fits the first moments of the marginal distributions.
1.4.2. Empirical Procedure

Under these two assumptions, we can estimate the OOI using a simple procedure that we
will now describe. The key idea of our method is to use the result of Lemma 1.1, that the
probability density \( f^i_j \) can be written as the ratio between the probability of observing a
match in the real distribution to its probability under a random assignment.

We start by expanding our data set of worker and job matches. We simulate data from
a distribution \( \tilde{f}(x,z) = f(x) \cdot f(z) \), where \( x \) and \( z \) are independent. This is done by
randomly sampling an observed worker and an observed job independently. We simulate
a total number of random matches equal to our original data size, such that the share of
real and simulated data is exactly one half. We define a binary variable \( Y \) that equals to
one whenever the match is ‘real’ (taken from the data) and zero whenever it is simulated.

We then estimate all our parameters using a logistic regression. We regress the binary
variable we constructed \( Y_k \) on the matched worker and job characteristics \((X_k, Z_k)\). Note
that, as a result of Lemma 1.1, and a simple Bayes rule, the match probability density \( f^i_j \) is
proportional to the ratio of observing this match in the real or simulated data, conditional
on the observed worker and job characteristics.

\[
\frac{P(Y_k = 1|x_k, z_k)}{P(Y_k = 0|x_k, z_k)} = \frac{f(x_k, z_k) \cdot P(Y_k = 1)}{f(x_k) \cdot f(z_k) \cdot P(Y_k = 0)} = f^i_j \cdot \text{const}
\]

Combining this result with Assumption 1.2 yields

\[
\log \frac{P(Y_k = 1|x_k, z_k)}{P(Y_k = 0|x_k, z_k)} = x_k A z_k + a(x_k) + b(z_k)
\]

We can estimate this equation using a logistic regression where we approximate \( a(x) \), \( b(z) \)
with linear functions. Under the assumptions this produces consistent estimates for
\( \hat{A}, \hat{a}(x_k), \hat{b}(z_k) \). We discuss the intermediate results from this estimation procedure, in
Section 1.7.2 where we analyze the underlying reasons for differences in the OOI.

We use the estimates from the logistic regression to estimate the probability density of every potential match. Specifically, we estimate the probability density of worker $i$ to work in job $j$ to be

$$
\hat{f}^i_j = \exp \left[ x_i \hat{A} z_j + \hat{a} (x_i) + \hat{b} (z_j) \right]
$$

We calculate this value for all possible worker-job combination in our data set. This simple functional form allows us to make this calculation directly and with minimum computational burden.

With these result in hand we can calculate the outside options index for every worker in our sample using Equation 1.8:

$$
\mathrm{OOI}_i = - \sum_j \hat{f}^i_j \log \hat{f}^i_j \quad (1.11)
$$

This yields a consistent estimate of the OOI, if both assumptions are correctly specified.

We verify that the OOI is robust to different choices of functional form. Instead of estimating it using the entropy index, we use the same probabilities we estimated in an HHI formula: $- \sum_j \hat{f}^2_0$. We find that the results are very similar. The correlation between the two indices is .62.

In Appendix A.1.3 we discuss the properties of this method, in the case where these assumptions do not hold. We show this method can be written as a GMM estimator, and discuss the moments that are being matched. We also show that if we increase the size of the simulated data, and fully saturate the functions $a$ and $b$, our method becomes equivalent to Dupuy and Galichon (2014).

---

23We normalize those estimated densities, such that $\sum_j \hat{f}^i_j = 1$. So effectively, we don’t use $\hat{a} (x_i)$. 

33
1.4.3. OOI Input: Job and Worker Characteristics

To estimate $f_j$ using Equation 1.10 we include three groups of variables: worker characteristics ($x$), job characteristics ($z$), and the geographical distance between workers and jobs. Information on wages is intentionally not used in any of these groups, to avoid a mechanical link between the OOI and wages.

**Worker Characteristics** $x$ We use $x$ to denote the variables that describe worker demographics and worker training. The demographic variables include workers’ gender, worker’s level of secondary education, an indicator for whether the worker is a citizen, and a quadratic in age. For training we use the occupation in which they undertook their apprenticeship. If we do not have information on a worker’s apprenticeship (e.g. if it occurred before our data begin in 1993), or if a worker did not complete an apprenticeship, we use their first occupation observed in the data, as long as this is at least ten years old.

**Job Characteristics** $z$ The job characteristics $z$ variables fall into three categories: (1) characteristics of establishments, (2) characteristics of employment contracts, and (3) characteristics of jobs. First, we take several establishment-specific variables directly from the establishment survey: size, sales and the share of females in management. We also use the first two principal components of each of the six categories of the establishment survey: business performance, investments, working hours, firm training, vocational training, and a general category. Table 1.2 shows the most weighted questions in each category. Second, we use several variables which relate to the structure of the employment contract: whether the job is part-time, whether the contract is fixed term, and whether the position was filled by a temporary agency.

Finally, to describe the characteristic of the job we use information on the occupation and
Table 1.2.: Most Weighted Question in PCA - Establishment 2014 Survey

<table>
<thead>
<tr>
<th>Name</th>
<th>N</th>
<th>Comp 1</th>
<th>Comp 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Performance</td>
<td>8,792</td>
<td>Member of chamber of industry</td>
<td>Profit category</td>
</tr>
<tr>
<td>Investment &amp; Innovation</td>
<td>8,792</td>
<td>IT investment</td>
<td>Total investment</td>
</tr>
<tr>
<td>Hours</td>
<td>8,792</td>
<td>Long leaves policy</td>
<td>Flextime</td>
</tr>
<tr>
<td>In-Company Training</td>
<td>8,792</td>
<td>Internal courses</td>
<td>Share workers in training</td>
</tr>
<tr>
<td>Vocational Training</td>
<td>8,792</td>
<td>Offer apprenticeship</td>
<td>Ability to fill</td>
</tr>
<tr>
<td>General</td>
<td>8,792</td>
<td>Family managed</td>
<td>Staff representation</td>
</tr>
</tbody>
</table>

This table shows the survey question that received the most weight in this principal component. We take the first two principal component from each survey category.

industry. Because it would not be feasible to include interactions between all of our industry and occupation codes, we use data from the BIBB to identify the characteristics associated with different industries and occupations. The BIBB survey contains modules on working hours, task type, requirements, physical conditions and mental conditions. For each 3-digit occupation and 2-digit industry, we include the first two principal components for each module. We use these to code both the occupation and industry that describe the job, and the training occupation that describes the worker. Appendix Table 1.3 shows the most weighted questions in each module. We also include occupation complexity, which codes occupations into four categories based on the type of activity they require: (1) simple, (2) technical (3) specialist and (4) complex. We use a total of 18 worker characteristics and 39 job characteristics.

Geographical Distance We include the geographical distance between workers and employers. For workers we use their last place of residence before taking the job. This

24These four categories usually reflect the type of qualification needed to perform the job, which ranges between none, vocational training, some tertiary degree and higher education. For instance, different occupations in nursing that fall under the same occupational coding (813) will be coded with different complexity, ranging between a nursing assistant, nurse, specialist nurse and general practitioner.

25Workers current place of residence is affected by their match. Their place of residence before taking the job better reflects the actual radius over which people are searching for jobs.
Table 1.3.: Most Weighted Question in PCA - BIBB

<table>
<thead>
<tr>
<th>Name</th>
<th>N</th>
<th>Comp 1</th>
<th>Comp 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours</td>
<td>11,021</td>
<td>Sundays and public holidays</td>
<td>hours per week like to work</td>
</tr>
<tr>
<td>Type of Task</td>
<td>15,035</td>
<td>responsibility for other people</td>
<td>Cleaning, waste, recycling</td>
</tr>
<tr>
<td>Requirements</td>
<td>10,904</td>
<td>Acute pressure &amp; deadlines</td>
<td>Highly specific Regulations</td>
</tr>
<tr>
<td>Physical</td>
<td>20,036</td>
<td>Oil, dirt, grease, grime</td>
<td>pathogens, bacteria</td>
</tr>
<tr>
<td>Mental</td>
<td>17,790</td>
<td>Support from colleagues</td>
<td>Often missing information</td>
</tr>
</tbody>
</table>

This table shows the survey question that received the most weight in this principal component. We take the first two principal component from each survey category.

distance could capture both the commuting, as well as the moving costs between places; empirically we cannot directly distinguish the two. Both locations are given at the district (kreis) level.

Figure 1.1 presents a map of the 402 districts in Germany. The size of the districts varies across the country and, importantly, it tends to be smaller in highly populated areas. In many cases, the major city is its own district, allowing us to separately identify the city center and the suburbs. Though not perfect, this coding allows us to get a reasonable approximation of commuting and moving patterns by workers. Appendix Table 1.4 shows the mean of the distance variable by gender and education groups. We find that the mean is 15.5 miles, but there is significant variation across groups.
Figure 1.1.: German Districts

Note: This map illustrates the 402 districts (kreis) in Germany.

We allow distance to have a non-linear effect on match probability that is different for each worker type. When we estimate Equation 1.10 we use a 4th degree polynomial of the distance between a worker’s lagged home district and their location of work to account...
Table 1.4.: Commuting Distance by Gender and Education

<table>
<thead>
<tr>
<th>Distance from Job (Miles)</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>15.5</td>
<td>41.9</td>
</tr>
<tr>
<td>Female</td>
<td>12.1</td>
<td>37.1</td>
</tr>
<tr>
<td>Male</td>
<td>17.4</td>
<td>44.3</td>
</tr>
<tr>
<td>Lower-Secondary</td>
<td>9.4</td>
<td>27.9</td>
</tr>
<tr>
<td>Intermediate-Secondary</td>
<td>11.4</td>
<td>34.4</td>
</tr>
<tr>
<td>Higher-Secondary</td>
<td>26.2</td>
<td>56.1</td>
</tr>
</tbody>
</table>

Values are mean distance in miles between workers previous place of residence and their job.

for the non-linear impact of distance.\textsuperscript{26} To account for heterogeneity in willingness to commute or move, we interact the polynomials in distance with all worker characteristics \( x \). This allows workers to be affected differently by distance, depending on their gender, education, age, citizenship and training. As we discuss in Section 1.7.2 this turns out to be the main driver of differences in outside options.

1.5. The Empirical Distribution of Outside Options

We next turn to describing the distribution of the OOI, and the characteristics of workers with better and worse options, as measured by it. We find that the OOI is higher for men, German citizens, city residents, more educated and more experienced workers. We also find that higher skill workers tend to be more specialized in their current industry, which narrows down their outside options.

\textsuperscript{26}We find that for more than 100 miles, the effect of distance is constant. This is consistent with the idea that individuals do not commute more than 100 miles; to switch to a job that is much further away, they have to move. This moving cost may not vary significantly with distance.
Figure 1.2: Distribution of Outside Option Index

Note: This figure plots the distribution of the outside options index as calculated for the population of German workers as of June 30th, 2014. The OOI was calculated using the procedure described in Section 1.4.2. LIAB sample weights are used to make the distribution representative of the German population.

Figure 1.2 plots the raw distribution of the OOI for every worker in our data. The mean of the distribution is $-4.85$. We can interpret the mean by considering the share $p$ of options a worker with this OOI would have if the probability density they worked at any given job was either $\frac{1}{p}$ or 0. A worker with an OOI of $-4.85$ would be found in a share $p = 0.8\%$ of jobs. The distribution is skewed, with a long left tail, indicating that there are many workers who are extremely concentrated. The standard deviation of the distribution is also quite sizable: .93. For comparison, duplicating the worker’s option set by generating an additional identical job for every job option they have would increase the OOI by only .69.
Figure 1.3.: OOI by Characteristics

Note: This figure plots the coefficients from a regression of OOI on education, gender, citizenship and a quadratic in age. The results are also presented on column 1 of Table 1.5. Confidence intervals are plotted at the 95% level. The lower axis shows raw OOI units, while the upper axis uses standard deviation units.
We estimate the following regression to decompose the average OOI by worker characteristics

\[ OOI = \beta_0 \text{Female} + \beta_1 \text{Education} + \beta_2 \text{Citizen} + \beta_3 \text{Age} + \beta_4 \text{Age}^2 + \epsilon \] (1.12)

Figure 1.3 plots the results. With controls, the average OOI for women is .237 units below that of men. The average OOI for German citizen is higher by .217 units. Assuming similar distributions across jobs, this would imply that male (German citizens) have 27% (24%) more options than women (non-citizens).

Options are also better for higher-educated workers. Lower-secondary (intermediate secondary) school workers’ options are on average .70 (.32) units lower than higher-secondary workers. This implies 101% (38%) more options assuming similar distribution across jobs.
**Figure 1.4.:** Cumulative Distribution of Outside Option Index by Gender

![Cumulative Distribution of Outside Option Index by Gender](image)

**Note:** This figure plots the cumulative distribution function of the outside options index by gender, as calculated for the population of German workers as of June 30th, 2014. The OOI was calculated using the procedure described in Section 1.4.2. LIAB sample weights are used to make the distribution representative of the German population.

Figure 1.4 shows that men have more options than women, not just on average, but across the entire distribution. The figure shows that the cumulative distribution function for men is shifted to the right. We cannot reject that the distribution for men stochastically dominates that of women.

We find an inverse U-shape relationship between the OOI and age. Figure 1.5 plots the mean OOI by age and shows that workers’ options tend to improve with age before flattening off at age thirty. Older workers (over 50) see declining values of options. As we discuss in Section 1.2.6, the OOI does not capture any dynamic considerations that are particularly likely to have a differential effect across ages. Accounting for this could po-
Figure 1.5.: OOI by Age

Note: This figure plots the mean OOI by age in the German population. LIAB sample weights are used to make the sample representative of the German population. Confidence intervals are plotted at the 95% level.

tentially change these results.

A large portion of the variation in options is driven by geographical variation in labor market size and density. The last category in Figure 1.3 shows a positive correlation between the district density and its OOI, controlling for other demographics. Figure 1.6 graphs the mean value of the OOI by German district. As the figure illustrates, workers in cities tend to have better options, as measured by the OOI. Workers near these cities, also appear to have better options. This result is robust for adding controls for worker demographics.

While most of our results indicate that high earning workers, such as workers with higher
Figure 1.6.: OOI Distribution by Region

Note: This figure plots the distribution of the outside options index by district (kreis) as calculated for the population of German workers as of June 30th, 2014. The OOI was calculated using the procedure described in Section 1.4.2. The value for each district is a weighted mean of the workers in this district, using the LIAB sample weights to make the distribution representative of the population in the district.
Note: This figure plots the mean residualized outside options index and log wages by training occupation as calculated for the population of German workers as of June 30th, 2014. The OOI was calculated using the procedure described in Section 1.4.2. Residuals for the OOI and log wages were taken from a regression on gender, a quadratic in age, education category, citizenship status and district of residence. Means are calculated using the LIAB sample weights to make the distribution representative of the population in the occupation. See Section 1.3.1 for exact definition of a training occupation.

education or city residents, tend to have more options, this relationship is reversed at the occupation level. This is because high-skilled workers tend to have more specialized skills, which are valued by a smaller number of employers. Controlling for all other observables, workers who completed their training (apprenticeship) in higher earning occupations tend to have lower options, as measured by the OOI (raw correlation equals -.022). Figure 1.7 plots each training occupation by their (residualized) log wages and OOI.
<table>
<thead>
<tr>
<th></th>
<th>Dep Var: Outside Option Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Female</td>
<td>-.237***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
</tr>
<tr>
<td>School</td>
<td>-.660***</td>
</tr>
<tr>
<td>Lower-Secondary</td>
<td>(0.013)</td>
</tr>
<tr>
<td>School</td>
<td>-.279***</td>
</tr>
<tr>
<td>Intermediate</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Non-Citizen</td>
<td>-.307***</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
</tr>
<tr>
<td>Age</td>
<td>.099***</td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
</tr>
<tr>
<td>Age^2</td>
<td>-.001***</td>
</tr>
<tr>
<td></td>
<td>(4e-05)</td>
</tr>
<tr>
<td>District</td>
<td>.112***</td>
</tr>
<tr>
<td>Density</td>
<td>(0.004)</td>
</tr>
<tr>
<td>Training Occ FE</td>
<td>X</td>
</tr>
<tr>
<td>District FE</td>
<td></td>
</tr>
<tr>
<td>Establishment FE</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.13</td>
</tr>
<tr>
<td>$N$</td>
<td>380,109</td>
</tr>
</tbody>
</table>

Notes: This table shows the results of a regression of OOI on basic demographics (Equation 1.12). The sample includes all workers employed on June 30th 2014. Sampling weights are used to make this a representative sample of the German population. Training occupation fixed effects are at the 3-digit levels.

We next look at which occupations drive the negative correlation. The upper-left corner of this figure, comprises high paying occupations with few relevant options: medical doctors, pilots, dentists. These are textbook examples of high-wage occupations with skills that cannot be easily transferred. While wages in these occupations are still high, our model predicts that, at least in partial equilibrium, if these workers were able to use their skills in more industries, their wages would have been even higher. The bottom right corner shows occupations like meter reader or car sales that have lower wages and more options. These are examples of low-wage occupations with highly general skills.
Table 1.5 presents coefficients from Equation 1.12, including additional controls for training occupation, district of residence and establishment. Controlling for training occupations (column 2) does not change the results significantly. However, adding controls for worker’s district of residence as well (column 3) reduces some of the education gap, and increases the gap between German citizens to non-citizens. This suggests that higher educated, and non-citizen workers are more concentrated in large cities where there are more job options, and therefore their OOI is lower once controlling for that. Controlling for establishments (column 4) yields results that are similar to the results with controls for districts and occupation, as workers in the same establishment tend to live closely. However, we find smaller gender differences in options within establishments.

1.5.1. Mass Layoffs

We next show that the OOI is able to predict the ease with which workers recover from a job separation. These separations force workers to move to their outside option. To identify exogenous separations we follow the prior literature in focusing on mass-layoffs.

We start by constructing a sample of workers who were involved in mass-layoff between 1993 and 2014, following the approach in Jacobson et al. (1993). We define a plant in our sample as undergoing a mass layoff if it has a decline in its workforce of at least thirty percent over the year. We consider only mass layoffs that occur in establishments with at least fifty workers. We restrict our analysis to workers who had been employed at the establishment for at least three years prior to the mass layoff and who are below the age of 55. This leaves us with a final sample of 13,681 workers from 583 distinct mass-layoffs.

The outcome variable we use is relative income: the ratio between current daily income and the last daily income before the layoff. Formally we define relative income as \( \bar{w}_t = \frac{w_t}{w_0} \), where \( t \) is months after the layoff, ranging from one to thirty-six. In case a worker is
### Table 1.6: Summary Stats for Mass-Layoff Workers by Treatment Status

<table>
<thead>
<tr>
<th></th>
<th>Above Median OOI</th>
<th>Below Median OOI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Female</td>
<td>.36</td>
<td>.48</td>
</tr>
<tr>
<td>Age</td>
<td>40.0</td>
<td>9.7</td>
</tr>
<tr>
<td>Higher-Secondary Education</td>
<td>.21</td>
<td>.41</td>
</tr>
<tr>
<td>Tenure in Establishment (days)</td>
<td>2316.3</td>
<td>1272.3</td>
</tr>
<tr>
<td>Daily Income</td>
<td>63.8</td>
<td>43.1</td>
</tr>
</tbody>
</table>

Note: This table shows the summary stats for workers that lost their jobs in a mass-layoff above and below the establishment median OOI. Mass layoffs are defined as an establishment with at least 50 workers that reduced its workforce by at least 30% in a given year. The sample includes only workers who have worked for at least three years before the layoff and are below the age of 55. We include monthly income for the 36 months following the separation.

Unemployed during this period their relative wage is set to zero. This choice of outcome variable takes out all productivity differences that can be captured with worker fixed effects and might be correlated with the OOI.

Figure 1.8 replicates the main result in Jacobson et al. (1993). Specifically, we look at

$$\tilde{w}_{i,t} = \beta_t + \psi_{j(i)}$$

where $\psi_{j(i)}$ is establishment fixed effect. We plot $\beta_t$, the mean relative income of workers each month, for three years after the layoff. We find that on average workers lose 80% of their income in the month following the layoff. Their income gradually returns to its previous value over the next three years.

We next look at the differences in recovery for workers with different value of OOI. Within each establishment in our sample, we divide the laid-off workers into two groups, based on whether they are above or below the establishment median of the outside options index. Table 1.6 shows summary statistics for the two groups.
Figure 1.8.: Relative Income Following Mass-Layoff

Note: This figure shows the relative income for workers who lost their jobs in mass-layoffs, for each month in the three years after the layoff. Relative income is defined as the current daily income in that month divided by the last daily income before the layoff. Mass layoffs are defined as an establishment with at least 50 workers that reduced its workforce by at least 30% in a given year. The sample includes only workers who have worked for at least three years before the layoff and are below the age of 55. The values are calculated using a regression of relative income on months after separation, with a fixed effect for every mass-layoff (Equation 1.13).
We then calculate

$$\bar{w}_{i,t} = \rho_t \text{High}_i + \delta_{j(i),t}$$  \hspace{1cm} (1.14)

where $\text{High}_i$ indicates above the establishment median OOI, and $\delta_{j(i),t}$ is establishment by month fixed effects. Figure 1.9 plots $\rho_t$, the difference in relative income between those two groups for each month in the three years after the layoff. Our point estimates show that workers with better options, as captured by the OOI, gain an additional 8 percent of their previous income during the first year after the layoff. The groups seem to converge throughout time and after three years there are no differences.

The lower relative income is driven by both longer search time, and lower wage after search. To show this we recalculate Equation 1.14, replacing the outcome variable with $e_{i,t}$ an indicator for being employed, which we define as having a positive wage

$$e_{i,t} = \bar{\rho}_t \text{High}_i + \delta_{j(i),t}^e$$  \hspace{1cm} (1.15)

Figure 1.10 plots $\bar{\rho}_t$, the differences in the share of employed workers on both groups. Our point estimate show that about 2% more people with higher OOI are working compared to the lower OOI. Therefore, the relative income in the new job must also be lower to explain an 8% difference between the groups.

We then repeat this analysis with a continuous measure of OOI, and a varying set of controls. We regress relative income at month $t$ on the OOI, with fixed-effects for establishment by month $\delta_{j(i),t}$.

$$\bar{w}_{i,t} = \lambda_t OOI_i + X_{it} + \delta_{j(i),t}$$  \hspace{1cm} (1.16)

We also repeat this analysis with additional worker controls $X_{it}$ including tenure, gender, age and education. The results are reported in Table 1.7. We find virtually the same patterns we found when divided by the median, for all choices of controls.
Figure 1.9.: Mass-Layoffs - Differences in Relative Income Between High/Low OOI Workers

Note: This figure shows the difference in relative income for workers with OOI above and below the establishment OOI. Relative income is defined as the current daily income in that month divided by the last daily income before the layoff. Mass layoffs are defined as an establishment with at least 50 workers that reduced its workforce by at least 30% in a given year. The sample includes only workers who have worked for at least three years before the layoff and are below the age of 55. The median OOI is calculated based on the pool of laid-off workers in a given establishment and year. The coefficients are taken from a regression of relative income on an indicator for above median OOI, interacted with indicator for each month after separation (plotted), with fixed effects for establishment × month (Equation 1.14).
Figure 1.10.: Mass-Layoffs - Differences in Search Time Between High/Low OOI Workers

Note: This figure shows the difference in employment for workers with OOI above and below the establishment OOI. Employment is defined as any income greater than zero. Mass layoffs are defined as an establishment with at least 50 workers that reduced its workforce by at least 30% in a given year. The sample includes only workers who have worked for at least three years before the layoff and are below the age of 55. The median OOI is calculated based on the pool of laid-off workers in a given establishment and year. The difference is calculated using a regression of employment on an indicator for above median OOI, interacted with indicator for each month after separation, with fixed effects for establishment×month (Equation 1.15).
Table 1.7.: Relative Income by OOI After Mass Layoff

<table>
<thead>
<tr>
<th>$OOI_i$ coefficient</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 months ($\lambda_3$)</td>
<td>.061**</td>
<td>.062**</td>
<td>.068**</td>
</tr>
<tr>
<td></td>
<td>(.029)</td>
<td>(.029)</td>
<td>(.031)</td>
</tr>
<tr>
<td>6 months ($\lambda_6$)</td>
<td>.068**</td>
<td>.069**</td>
<td>.082**</td>
</tr>
<tr>
<td></td>
<td>(.030)</td>
<td>(.030)</td>
<td>(.033)</td>
</tr>
<tr>
<td>12 months ($\lambda_{12}$)</td>
<td>.061*</td>
<td>.064*</td>
<td>.079**</td>
</tr>
<tr>
<td></td>
<td>(.034)</td>
<td>(.034)</td>
<td>(.038)</td>
</tr>
<tr>
<td>24 months ($\lambda_{24}$)</td>
<td>.033</td>
<td>.039</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td>(.042)</td>
<td>(.042)</td>
<td>(.048)</td>
</tr>
<tr>
<td>Mass-Layoff × Month FE</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Tenure</td>
<td>Y</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>Y</td>
</tr>
</tbody>
</table>

No. of Observations 558,686 558,686 558,686  
No. of Workers 13,707 13,707 13,707  
No. of Mass-Layoffs × Month 26,561 26,561 26,561

Note: This table shows the results of regressing relative income on OOI for workers that lost their jobs in a mass-layoff, for different times after the separation. Relative income is defined as the current daily income in that month divided by the last daily income before the layoff. Mass layoffs are defined as an establishment with at least 50 workers that reduced its workforce by at least 30% in a given year. The sample includes only workers who have worked for at least three years before the layoff and are below the age of 55. We include monthly income for the 36 months following the separation. The regression is based on Equation 1.16. Tenure includes a quadratic polynomial for days at the previous establishment. Age includes a quadratic polynomial. Education is a categorical variable for the type of secondary education (see section 1.3.2 for details).
Our findings suggest that the OOI may have an impact on additional dynamic aspects of the labor market that are not captured in our static-model. The OOI quantifies the number of options workers have. It is therefore not surprising that workers with more options (higher OOI) are able to find a new job more quickly. However, the effect on relative wage is less obvious, as the OOI is likely to already affect income before the layoff. One interpretation of our findings is that workers with higher OOI face a labor market that is closer to perfectly competitive, such that the impact of a single employer on equilibrium wages is negligible, and not affected by their exit.27

1.6. Effect on Wages

In this section we use two different sources of quasi-random variation in options to estimate the elasticity between the OOI and wages. The first (Section 1.6.1) focuses on the introduction of high-speed commuter rail stations in small German towns, and the second (Section 1.6.2) uses a shift-share (“Bartik”) instrument. These methods yield semi-elasticities between .17-.32 .

Estimating this elasticity using instruments allows us to translate differences in OOI to differences in wages, even if the model is misspecified. In our model, the effect of the OOI on wages is determined by the parameter $\alpha$ (Theorems 1.1, 1.2). Our model only implies that this parameter is non-negative and that in a perfectly competitive labor market, this parameter is zero. Since we do not use any of the model assumptions to estimate this parameter, the results capture the elasticity between the OOI and wages, even if the model is misspecified. This exercise can also be seen as a test for whether the OOI has an impact

27This intuition can be captured with the predicted effect of a plant closure on the OOI. Removing a job with low probability have only a small effect on the OOI. Since workers with higher OOI, have on average a lower probability to be in every job, their OOI is less affected from the destruction of only a few jobs. Therefore their wage will be less affected as well.
Table 1.8.: Correlation Between OOI and log wage - OLS Results

<table>
<thead>
<tr>
<th></th>
<th>Dep.Var: log wage&lt;sub&gt;i&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOI</td>
<td>.107*** .027*** -.010*</td>
</tr>
<tr>
<td></td>
<td>(.005) (.005) (.005)</td>
</tr>
<tr>
<td>Demographics</td>
<td>X</td>
</tr>
<tr>
<td>District FE</td>
<td>X</td>
</tr>
<tr>
<td>N</td>
<td>378,776</td>
</tr>
</tbody>
</table>

Note: Demographics include gender, education group, a quadratic in age and citizenship status. The sample includes all workers employed on June 30th 2014. Sampling weights are used to make this a representative sample of the German population.

on wages.

Identifying α, the relationship between options and wages is challenging for two reasons. First, the OOI estimator is a function of the observed characteristics X<sub>i</sub>. These observables are likely to also capture differences in productivity, thus creating a problem of omitted variable bias. Second, the OOI is estimated with a potentially large amount of noise, which would create an attenuation bias, especially when adding several controls. Because of those reasons, the OLS estimates of this parameter (Table 1.8), strongly depend on the set of controls we use. Adding more controls shrinks the results towards zero, and can even affect the sign of the coefficient. In order to cope with both issues we use two sources of quasi-random variation in workers’ options that do not affect their productivity.

1.6.1. High-Speed Commuter Rail

We leverage the expansion of the high-speed rail network in Germany as an exogenous shock to workers’ outside options, following prior work by Heuermann and Schmieder.

---

28Our findings in Section 1.5 suggest that the bias could go both ways, therefore the sign of the OLS coefficient could be both positive and negative. The OOI is correlated with high experience, or high education, but also with lower wage occupations, potentially because high-skill occupations tend to be more specific.
High-speed trains were first introduced in Germany in 1991-1998. During that period, stations were placed in major cities. We focus on the second wave of expansion, which began in 1999. During this wave, new stations were added in cities along existing routes. Cities were chosen for new stations after the routes already existed, and mostly on the basis of political considerations, and not labor market factors. As a result, towns as small as 12,000 residents were connected to the train network. Heuermann and Schmieder (2018) show that this increase in infrastructure led to an increase in commuting probability. Figure 1.11 shows the map of districts that got stations in the two waves of installation.

29Daniel Heuermann and Johannes Schmieder generously provided the train data for our use in this project.
Note: This figure shows the locations of ICE train stations by districts. The first wave includes all stations that were opened pre-1999. The second wave includes all stations that were opened post-1999.

There are several threats to identification that should be considered. One potential concern is that the cities which received new train stations were selected on the basis of ex-
pected increase in productivity. Institutional details and prior research suggest that this is unlikely. Another concern is that the new infrastructure can also be used for the transportation of goods. This would impact the workers’ wages through their productivity. However, Heuermann and Schmieder showed that the introduction of these stations had no effect on the product market, as the trains were only used to transport passengers. One remaining concern is that the trains also had a similar effect on employers in those small towns, by allowing them to recruit workers from major German cities. We do not have detailed enough data on the employers to reject this possibility.

We supplement our data with train schedules in the years 1999 and 2012, before and after the installment of the second-wave stations. From this data we construct an indicator variable for every match for whether the worker can use a direct high-speed line to get to this job. We add this variable for our estimation of the match probabilities $f_{ij}$. Therefore, we allow the match probability to depend on whether there’s a high-speed line connecting the worker and the employer. To allow for heterogeneity between workers in their demand for trains, we interact this variable with all worker characteristic $X_i$. We then use these probabilities to estimate the OOI, as explained in Section 1.4.2.

The treatment group consists of all workers who, in 1999, lived in districts that got high-speed connections during the second-wave expansion (1999-2012). The control group consists of all workers who, in 1999, lived in districts that never received stations. Since the major cities were connected in the first wave, they are effectively excluded in this analysis. We then follow the same workers to the year 2012, regardless of where they live. We match workers from the treatment and the control group based on their gender, age, citizenship, education level, training occupation, state (Bundesländer) and lagged income using nearest-neighbor matching with replacement. Table 1.9 presents a balance table

---

30See Heuermann and Schmieder (2018) for the full list of criteria that were used for location choices.
31We require the match to be exact on gender, education, state and 2-digit occupation.
Table 1.9: Balance Table - High Speed Train

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>log wage (1993)</td>
<td>3.26</td>
<td>2.49</td>
</tr>
<tr>
<td>log wage (1999)</td>
<td>4.25</td>
<td>.62</td>
</tr>
<tr>
<td>Female</td>
<td>.356</td>
<td>.479</td>
</tr>
<tr>
<td>Age</td>
<td>36.4</td>
<td>6.8</td>
</tr>
<tr>
<td>Citizen</td>
<td>.995</td>
<td>.073</td>
</tr>
<tr>
<td>Low-Secondary</td>
<td>.257</td>
<td>.437</td>
</tr>
<tr>
<td>Intermediate-Secondary</td>
<td>.508</td>
<td>.500</td>
</tr>
<tr>
<td>High-Secondary</td>
<td>.235</td>
<td>.424</td>
</tr>
<tr>
<td>N</td>
<td>37,695</td>
<td></td>
</tr>
</tbody>
</table>

Note: This table shows the summary stats for workers used to estimate the impact of high-speed trains on OOI and log wages. Treated group includes workers who lived in districts in which a new station was introduced between 1999-2012. Control group was chosen from a pool of workers living in districts that never got a station. Control workers were chosen through nearest-neighbor matching with replacement on gender, age, citizenship, education level, training occupation, state (Bundesländer) and lagged income. We require the match to be exact on gender, education, state and 2-digit occupation.

We estimate the following system of equations:

\[
\Delta_{2012}^{1999} \log w_{im} = \alpha \Delta_{1999}^{2012} OOI_i + \mu_m + \nu_{im} \\
\Delta_{1999}^{2012} OOI_{im} = \delta \text{Treat}_i + \lambda_m + \epsilon_{im}
\]

where \( \text{Treat}_i \) is an indicator for living in a treated district in 1999 and \( \mu_m, \lambda_m \) are match fixed-effect. Because this is a binary instrument, \( \hat{\alpha} \) collapses to a Wald estimator

\[
\hat{\alpha} = \frac{\Delta_{1999}^{2012} \log w_{treatment} - \Delta_{1999}^{2012} \log w_{control}}{\Delta_{1999}^{2012} OOI_{treatment} - \Delta_{1999}^{2012} OOI_{control}}
\] (1.17)

where the average is taken over matched pairs. We develop a procedure that calculates
Table 1.10: Impact of Express Trains on Options and Wages

<table>
<thead>
<tr>
<th></th>
<th>First-Stage 1999-2012 (1)</th>
<th>Reduced-Form 1999-2012 (2)</th>
<th>2SLS 1999-2012 (3)</th>
<th>Reduced-Form 1993-1999 (4)</th>
<th>OLS 1999-2012 (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.073***</td>
<td>.024***</td>
<td>.324***</td>
<td>.002</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.004)</td>
<td>(.048)</td>
<td>(.002)</td>
<td>(.007)</td>
</tr>
</tbody>
</table>

Number of observations: 143,313
Number of treated observations: 37,695

Notes: This table shows the results of the impact of express trains on outside options, and wages. Columns 1-4 use nearest-neighbor matching with replacement. Matching is done exactly on gender, education group, citizenship status, state and 2-digit training occupation and continuously on age, and PCA components for training occupation (the third digit). The outcome variables are change in OOI 1999-2012 (column 1), change in log income 1999-2012 (columns 2,3,5) and change in log income 1993-1999 (column 4). Standard errors in matching are calculated using Abadie and Imbens (2006). 2SLS estimator is the division of the estimates in column 1 and 2 (Equation 1.17). Standard errors in column 3 are calculated using a method building on Abadie and Imbens (2006) (see Appendix A.1.4 for details). OLS (column 5) estimates the regression of log wages on OOI with match fixed effects. Observations from the control group that appear in multiple matches also appear multiple times in the OLS. Standard errors are clustered for workers with the same variables we match on exactly to account for the replacement (see Appendix A.1.4 for details).


Table 1.10 shows the main result: an elasticity of .32 between options and wages. Column 1 shows that the OOI increased by .07 in treated districts following the introduction of the new stations. The reduced form results in column 2 suggest an increase of about 2.5% increase in income in the treated districts. Combining both estimates into a 2SLS estimator in column 3 yields a semi-elasticity of approximately .32 between our measure of outside options and wages. Column 4 shows that our matching process worked: there are no pre-trends. Our OLS results in column 5 show a precise zero. This is likely to be driven by an attenuation bias that is amplified substantially when using first differences
We next verify that our effect is driven by workers who are more likely to use the train. The high-speed commuter rail is a fairly expensive commuting option. As a result, the introduction of train stations should primarily affect high-income workers. We break our sample into three education groups, which we use as a proxy for potential income. Figure 1.12 plots the first stage and the reduced form results, together with our point 2SLS estimates for each group. We find a higher first stage for workers with higher education. These are the workers we would expect to use the train the most. The reduced form is also higher for the more educated workers, though the estimate is imprecise. We cannot rule out a zero effect on the low education group. The two-stage least squares estimates are similar for all three groups, and we cannot rule out homogeneous effects by education group.

1.6.2. Shift-Share (“Bartik”)

We next use a standard shift-share (“Bartik”) instrument to estimate the elasticity between wages and options. Though this exercise gives a lower point estimate of .17, we cannot reject that the elasticity we estimate is identical to the one estimated in the prior section. The idea behind this strategy is to compare workers who work in the same industry, but who have different outside options, because they reside in different parts of the country with different industry mixes. Some workers happen to live near industries that are growing, while others happen to live near industries that are contracting. Because local

---

32 Duncan and Holmlund (1983) show that this depends on the level of autocorrelation between the measurement errors, and true signal. Since worker observables tend to be constant while the measurement error might change between years, we expect attenuation bias to be much stronger in first difference.

33 For example, a round-trip between Montabaur and Frankfurt takes 45 minutes each way and costs 60 Euros.

34 These shocks were used in several papers including Bartik (1991); Blanchard and Katz (1992); Card (2001); Autor et al. (2013). It was used specifically for the context of a shock to outside options by Beaudry et al. (2012).
Figure 1.12: Impact of Express Trains by Schooling Level

Note: This figure plots the first-stage and reduced-form results for three education groups, and their combination. First stage is the treatment effect on OOI. Reduced form is the treatment effect on log wages. Both were calculated using nearest-neighbor matching with replacement. Treatment is defined as workers that in 1999 lived in districts that got ICE stations post-1999. The control group includes workers that in 1999 lived in districts that never got ICE stations. Matching is done exactly on gender, education group, citizenship status, state and 2-digit training occupation and continuously on age, and PCA components for training occupation. Confidence intervals are at the 95% level, and are calculated based on standard errors derived from a method by Abadie and Imbens (2006). The black line represents the 2SLS point estimate for the entire sample.
growth of certain industries may be due to the impact of local productivity shocks, we use national industry trends as an instrument.

The instrument is a weighted average of national industry growth, weighted by the initial share of each industry in the region. Formally, we define

\[
B_r = \sum_j s_{jr}^{04} \times \hat{g}_j
\]

where \(s_{jr}^{04}\) is the share of employed workers in region \(r\), working at industry \(j\) in the base year (2004) and \(\hat{g}_j\) is the national employment growth of industry \(j\). Regions are defined by the administrative regions ("Regierungsbezirke") in Germany, the statistical unit which is closest to a commuting zone.\(^{35}\) Industries are defined at the 3-digit level.

To estimate the national growth of different industries, controlling for region-wide shocks, we regress the change in employment in industry \(j\) in region \(r\) between 2004 and 2014 on industry and region fixed effects:\(^{36}\)

\[
\Delta_{14}^{04} \log E_{jr} = g_j + g_r + \varepsilon_{jr}
\]

By construction, the estimator of \(\hat{g}_j\) is not driven by regional trends captured in \(\hat{g}_r\). We use the weighted average of the industry fixed effects \(\hat{g}_j\) by initial industry shares \(s_{jr}^{04}\) to calculate \(B_r\). This construction verifies that \(B_r\) is not driven by local employment shocks in this region, or even in nearby regions.\(^{37}\)

---

\(^{35}\)We take all 39 regions based on NUTS2 level coding of the European Union. This includes historical administrative regions that have been disbanded. Results are robust to the definition of a region and hold for the NUTS3 level (district) as well.

\(^{36}\)We make a Bayesian correction of uniform prior by adding one observation in each industry, region and year combination.

\(^{37}\)This is different from a leave-one-out estimate, that might still be driven by local shocks in nearby regions.
Table 1.11.: Effect of OOI on Wages Using Shift-Share (Bartik) Instrument

<table>
<thead>
<tr>
<th></th>
<th>First-Stage (1)</th>
<th>Reduced-Form (2)</th>
<th>2SLS (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.622***</td>
<td>.106***</td>
<td>.170***</td>
<td></td>
</tr>
<tr>
<td>(.241)</td>
<td>(.056)</td>
<td>(.064)</td>
<td></td>
</tr>
<tr>
<td>Industry FE</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>N</td>
<td>408,792</td>
<td>408,792</td>
<td>408,792</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>38</td>
<td>38</td>
<td>38</td>
</tr>
</tbody>
</table>

Notes: This table shows the results of the impact of a shift-share instrument (Bartik) on outside options, and wages (column 1 and 2). This captures the effect of changes in OOI on changes in log wages between 2004-2014, when we instrument for the changes in OOI with the shift-share instrument (column 3). The instrument is constructed from an average of a 3-digit industry national employment growth weighted by the initial share of every industry in a region (see Section 1.6.2). The outcome variables are the change in OOI 2004-2014 (column 1) and change in log daily wages (columns 2 and 3). All columns control for industry (in 2004) and age. Standard errors are clustered within the unit of treatment, which is regions.

We estimate the following system of equations

\[
\begin{align*}
\Delta_{04}^{14} \log w_{ijr} &= \alpha \Delta_{04}^{14} OOI_{ijr} + \beta \Delta_{04}^{14} X_{ijr} + \text{Ind}_{04}^{04} + v_{ijr} \\
\Delta_{04}^{14} OOI_{ijr} &= \gamma B_r + \delta \Delta_{04}^{14} X_{ijr} + \text{Ind}_{04}^{04} + \epsilon_{ijr}
\end{align*}
\]

where we control for the industry of the worker in the beginning of the period (2004). We cluster standard errors at the level of the treatment, which is the region. The parameter of interest is \( \alpha \), the elasticity of wages with respect to options.

Table 1.11 presents the main results. Columns 1 and 2 show the first-stage and reduced form results. A 10% higher employment in other industries, which is about .1 increase in the instrument, translates to approximately 6% more relevant options, and 1% increase in wages. Combining both estimates yields a semi-elasticity of .17: a 10% increase in relevant options leads to a 1.7% increase in wages.

The identifying assumption is that growing industries are not systematically located in
regions where wages are growing for other reasons (Borusyak et al., 2018). One way the assumption could be violated is if there are productivity spillovers. Workers that live near industries that are growing, may enjoy a local demand shock for their production due to the positive income effect on workers in that region. This could generate a wage increase, that is not driven by the improvement in their outside options. This is particularly a concern for workers who are producing non-tradable goods, whose productivity is set by local demand.

We address this concern by showing the results hold for workers in exporting industries, which are less likely to be affected by local demand shocks. We use information from the establishment survey to calculate the export share of each industry.\(^{38}\) We divide our data into three groups based on the export share of the industry where the worker worked in 2004. Table 1.12 shows the results for each of the groups. We find a large and statistically significant elasticity between options and wages even among workers in industries with the highest exporting share. Column 1 indicates that in response to a 10% increase in OOI, workers in these industries see their wages rise by 1%. This elasticity is somewhat lower than that in our baseline results (.10 versus .17). However, we cannot reject that they are equal.\(^{39}\)

We next examine heterogeneity across gender and the three education groups. We estimate Equations 1.18 separately for each group. Figure 1.13 plots the results for all groups, as well as the full population. While splitting the sample increases the size of the confidence intervals, the point estimates are quite close. This suggests that using the same

---

\(^{38}\)This is a lower bound for demand from outside the region, as it does not include sales to other regions in Germany, which we cannot see in our data. We calculate the mean industry level since we don’t have the share of sales from export for all employers, but only a representative sample.

\(^{39}\)Beaudry et al. (2012) find similar results when dividing the data into tradable and non-tradable industries, based on their geographical spread. They argue that non-tradable industries are geographically spread across different regions, while tradable goods could be concentrated in specific regions. They also address additional potential threats to the identification assumption and show they don’t seem to have a significant effect on the result.
Table 1.12.: Shift-Share (Bartik) Results by Exporting Share of Sales

<table>
<thead>
<tr>
<th>Export</th>
<th>33% &lt; Export</th>
<th>≥ 1%</th>
<th>1% &gt; Export</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOI</td>
<td>.105**</td>
<td>.593**</td>
<td>.132</td>
</tr>
<tr>
<td></td>
<td>(.052)</td>
<td>(.266)</td>
<td>(.141)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Industry FE</th>
<th>X</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>119,645</td>
<td>146,217</td>
<td>142,930</td>
</tr>
</tbody>
</table>

Notes: This table shows the results of the impact of OOI on wages, instrumented with a shift share instrument, calculated separately by share of export in the industry. Share of export is calculated for every 3-digit industry based on the establishment panel survey in 2014. The sample is split based on the worker industry in 2004. Outcome variable is change in log wages between 2004-2014. The dependent variable is change in OOI between 2004-2014. The instrument is constructed from an average of a 3-digit industry national employment growth weighted by the initial share of every industry in a region (see Section 1.6.2). All columns control for industry (in 2004), and age. Standard errors are clustered within the unit of treatment, which is regions.

value for \( \alpha \) for all groups is a reasonable approximation.

We next use this setting to decompose the different effect of access to more options into impacts for job stayers and movers. Because the choice of whether to move is endogenous, we view this as a decomposition exercise. We interact the changes in OOI with an indicator variable for whether a worker stayed at their establishments during this period. The results are shown in Table 1.13. As our model predicts, we find that the effect on stayers is smaller. This is possibly because they only benefit through an improvement in their outside options. The larger effect on movers is consistent with an additional improvement in match quality.

While the elasticity we estimate in this exercise is lower from the one we estimated using the fast commuter rails, their difference is not statistically different from zero. Figure 1.14 compares our results in this section to the elasticity we estimated using the introduction of high-speed commuter rails. The fact that we found elasticities of a similar magnitude by using two distinct sources of variation suggests that this range of estimates is a reasonable
Note: Every category displays the estimate for coefficient $\hat{\alpha}$ from Equation 1.18, run separately for each education group or gender (blue), and for the entire population (red). This captures the effect of changes in OOI on changes in log wages between 2004-2014, when we instrument for the changes in OOI with the shift-share instrument. The instrument is constructed from an average of a 3-digit industry national employment growth weighted by the initial share of every industry in a region (see Section 1.6.2). Standard errors are clustered within the unit of treatment, which is regions. Confidence intervals are at the 95% level.
Table 1.13: Shift-Share (Bartik) Results by Stayers and Movers

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$OOI$</td>
<td>.170***</td>
<td>.257***</td>
</tr>
<tr>
<td></td>
<td>(.064)</td>
<td>(.092)</td>
</tr>
<tr>
<td>$OOI \times \text{Stay}$</td>
<td>-.159***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(.062)</td>
<td></td>
</tr>
<tr>
<td>Industry FE</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>$N$</td>
<td>408,792</td>
<td>408,792</td>
</tr>
</tbody>
</table>

Notes: This table shows the results of the impact of $OOI$ on wages, instrumented with a shift share instrument, interacted with whether a worker stayed at the same establishment. Outcome variable is change in log wages between 2004-2014. The dependent variable is change in $OOI$ between 2004-2014. The instrument is constructed from an average of a 3-digit industry national employment growth weighted by the initial share of every industry in a region (see Section 1.6.2). The indicator for stay is 1 if the worker works at the same establishment on June 30th of both 2004 and 2014. All columns control for industry (in 2004), and age. Standard errors are clustered within the unit of treatment, which is regions.

benchmark for the value of $\alpha$.

1.7. Effect on Wage Inequality

In this section, we combine our estimates on the distribution of the $OOI$, with our estimates of the $OOI$-wage elasticity, to assess the overall effect of options on the wage distribution. We then examine which covariates drive differences in options. We find that equalizing workers’ ability to commute or move would eliminate the gender gap in $OOI$, and would reverse the sign of the $OOI$ gap by education.
Figure 1.14: Estimates of Elasticity Between OOI and Wages

Note: This figure compares the elasticity between OOI and wages from the two different sources of quasi-random variations that we used. Train includes the results for parameter \( \hat{\alpha} \) estimated using the introduction of high-speed commuter rails (see Section 1.6.1 and notes for Figure 1.12 for more details). Shift-Share uses an instrument based on national industry employment trends (see Section 1.6.2 and notes for Figure 1.13 for more details). Confidence intervals are at the 95% level. The difference between the point estimates is .156 (.081).
1.7.1. Overall Impact on the Wage Distribution

We examine what portion of between-group wage inequality can be attributed to difference in OOI. We first estimate a Mincer equation

$$\log w_i = \beta_0 X_i + \epsilon_i$$

(1.19)

where $X_i$ includes indicators for each education group, a quadratic in age, gender, citizenship status, log district density and an indicator for part-time job. Since wages are top-coded we use a Tobit model to estimate $\hat{\beta}_0$. We then add the OOI to the set of dependent variables, with a fixed coefficient.

$$\log w_i = \hat{\alpha}_{\text{OOI}} + \beta_1 X_i + \epsilon_i$$

We use $\hat{\alpha} = .26$ which is the average of the two point estimates we derived in Section 1.6 from the two quasi-random sources of variation. $\hat{\beta}_0$ captures the overall gaps in wages between these demographic groups, $\hat{\beta}_1$ is the remaining gaps that are driven by factors other than the OOI, and $\hat{\beta}_0 - \hat{\beta}_1$ is the part that can be attributed to the differences in OOI.

Figure 1.15 shows the main results. The full bars display the gaps that we estimated ($\hat{\beta}_0$), where every bar is the wage premium for this group members. For instance, the premium for being a male (the gender gap) is $.19$ log units in Germany. The portion that can be attributed to the OOI ($\hat{\beta}_0 - \hat{\beta}_1$) is colored in red, while the remaining gap ($\hat{\beta}_1$) is left in blue.

The OOI explains significant portions of several German wage gaps. When we add OOI to the regression, the gender gap is cut by $.06$ log units (30% of the overall gap). This is driven by the .23 gender gap in OOI we found in Table 1.5, multiplied by $\hat{\alpha}$. Our results also indicate that 88% of the gap between German citizens to non-citizens (.08 log units)
Figure 1.15.: Overall Effect on Wage Inequality

Note: Every bar in this plot is the coefficient on the corresponding category in a regression of log wages on Male, Citizen, indicator for secondary-education category, a quadratic in age, district density and an indicator for part-time job. The blue portion of the bars (remaining gap) is the coefficient from the same regression, controlling for the OOI with a coefficient fixed to .26, which was estimated with the two quasi-random variations. The part in red (explained gap) is the difference between the two coefficients. The reference workers is a female, non-citizen, with intermediate secondary education and 18 years old.
can be attributed to differences in options. The wage difference between high-level and intermediate-level secondary schooling is cut by .07 log units, which is about 25% of the initial gap. Our results also attribute 39% of the return to experience at age 18 in access to options. Table 1.14 shows these results numerically in columns (3) and (4), as well as results from a winsorized OLS in columns (1) and (2).

1.7.2. Explaining Differences in Outside Options - The Role of Commuting Costs

We next examine which factors impact workers’ options. We find that differences in commuting costs seems to be particularly important, especially in its effect on the gender gap and return to higher education.

We start by examining the impact of different variables on the probabilities of observing a match. We analyze our results from the estimation of matrix $A$ defined in Assumption 1.2, which we estimated using a logistic regression. This matrix is also the cross-derivative of match quality $\tau$ (Equation 1.9). Table 1.15 shows the top absolute values of $\hat{A}$, when variables are standardized so the results are not affected by specific units.\footnote{Online appendix table A1 shows the full standardized results for $A$; online table A2 shows the raw results.}

These results indicate that the most important factor in determining match quality is commuting and moving costs. Distance has the largest standardized coefficient in absolute terms (-4.15). While distance to a job is an important factor for all workers, it is particularly important for female workers, for less-educated workers, and for non-German citizens. Appendix Table 1.16 presents the raw coefficient on distance for different worker characteristics. This coefficient is the effect of an additional mile on the log probability of a match, at mile zero. For our baseline group, forty year old males citizens from higher secondary schools, the coefficient is -.141. The interaction with female is -.024, so women
Table 1.14.: Mincer Equation with OOI

<table>
<thead>
<tr>
<th></th>
<th>OLS (1)</th>
<th>OLS (2)</th>
<th>Tobit (3)</th>
<th>Tobit (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOI (fixed)</td>
<td>-.269</td>
<td>-.269</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-.171***</td>
<td>-.111***</td>
<td>-.195***</td>
<td>-.137***</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>School Lower-Secondary</td>
<td>-.351***</td>
<td>-.174***</td>
<td>-.404***</td>
<td>-.230***</td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>School Intermediate</td>
<td>-.245***</td>
<td>-.170***</td>
<td>-.289***</td>
<td>-.217***</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.010)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>German</td>
<td>.089**</td>
<td>.007</td>
<td>.093***</td>
<td>.011</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
<td>(0.034)</td>
<td>(0.032)</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Age</td>
<td>.057***</td>
<td>.030***</td>
<td>.061***</td>
<td>.035***</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.004)</td>
<td>(0.003)</td>
<td>(0.004)</td>
</tr>
<tr>
<td>Age^2 (x 10^-3)</td>
<td>-.573***</td>
<td>-.261***</td>
<td>-.608***</td>
<td>-.297***</td>
</tr>
<tr>
<td></td>
<td>(0.039)</td>
<td>(0.042)</td>
<td>(0.040)</td>
<td>(0.042)</td>
</tr>
<tr>
<td>District</td>
<td>0.022***</td>
<td>-0.008**</td>
<td>0.023</td>
<td>-0.008**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
<td>(0.003)</td>
</tr>
<tr>
<td>Density (log)</td>
<td>-.913***</td>
<td>-.905***</td>
<td>-.928***</td>
<td>-.921***</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.015)</td>
<td>(0.015)</td>
<td>(0.015)</td>
</tr>
<tr>
<td>Part-Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>378,776</td>
<td>378,776</td>
<td>378,776</td>
<td>378,776</td>
</tr>
</tbody>
</table>

Notes: This table shows the results from a regression of log wages on demographics and OOI. The coefficient of the OOI is fixed to be its point estimate from the 2SLS estimate based on the high-speed commuter rail introduction (Table 1.10). A Tobit model is used in Columns 3-4 to account for top coding of daily income at 195 Euros per day. OLS results use winsorized log income. Sampling weights are used to make this a representative sample of the German population.
Table 1.15.: Top Standardized Values of $A$

<table>
<thead>
<tr>
<th>Variable ($X$)</th>
<th>Variable ($Z$)</th>
<th>$A_{xz}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance</td>
<td></td>
<td>-.415</td>
</tr>
<tr>
<td>Train Occ - Physical Cond. 1</td>
<td>Occ - Physical Cond. 1</td>
<td>1.477</td>
</tr>
<tr>
<td>Train Occ - Task Type 2</td>
<td>Occ - Task Type 2</td>
<td>1.077</td>
</tr>
<tr>
<td>Train Occ - Task Type 2</td>
<td>Occ - Physical Cond. 1</td>
<td>-.93</td>
</tr>
<tr>
<td>Train Occ - Physical Cond. 1</td>
<td>Occ - Task Type 2</td>
<td>-.82</td>
</tr>
<tr>
<td>Lower Secondary Education</td>
<td>Distance</td>
<td>-.74</td>
</tr>
<tr>
<td>Intermediate Education</td>
<td>Distance</td>
<td>-.61</td>
</tr>
<tr>
<td>Train Occ - Contract 2</td>
<td>Occ - Contract 2</td>
<td>.56</td>
</tr>
<tr>
<td>Train Occ - Task Type 1</td>
<td>Occ - Task Type 1</td>
<td>.55</td>
</tr>
<tr>
<td>Lower/Intermediate Education</td>
<td>Distance</td>
<td>-.54</td>
</tr>
</tbody>
</table>

Results from logistic regression for dummy variable on real vs. simulated match, on interaction of worker and job characteristics (Equation 1.10). Results are standardized, such that each variable has standard deviation of 1.

are 17% more sensitive to distance than the baseline group. Lower educated workers are significantly more sensitive to distance (coefficient -.037). Non-German citizens seem to be more sensitive than citizens (coefficient -.019). Finally, workers at first become less sensitive to distance with age, but this is a concave function that reaches its maximum at age 42.

By simulating counterfactuals from the underlying model, we can quantify the overall effect of differences in commuting and moving costs on wages through their effect on the OOI. We estimate the wage gain for every worker, if they had the minimal commuting/moving costs. Based on our estimation, these are the costs of a 40 year old, high-educated, male citizen. We generate a matrix $\tilde{A}$ where the coefficients on distance is set to this minimum level for all workers. We then simulate the probabilities $\tilde{f}_j^i$ using this matrix, calculate the $\tilde{O}\tilde{O}I_i$ and translate it to $\tilde{\log} w_i$ using $\hat{\alpha}$. This counterfactual should be thought of as changing only a zero measure number of workers each time, and keep-
Table 1.16: Distance Coefficient by Demographics

<table>
<thead>
<tr>
<th>Category</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>-.141</td>
</tr>
<tr>
<td>Female</td>
<td>-.024</td>
</tr>
<tr>
<td>Non-Citizen</td>
<td>-.019</td>
</tr>
<tr>
<td>Lower-Secondary</td>
<td>-.037</td>
</tr>
<tr>
<td>Intermediate-Secondary</td>
<td>-.012</td>
</tr>
<tr>
<td>Age</td>
<td>.002</td>
</tr>
<tr>
<td>Age^2 (\times 10^{-3})</td>
<td>-.026</td>
</tr>
</tbody>
</table>

Results from logistic regression for dummy variable on real vs. simulated match, on interaction of worker and job characteristics (Equation 1.10). Baseline category is a forty years old high-secondary male.

We compare the differential gains from this exercise to assess the importance of commute in generating wage gaps.

We run a regression of the counterfactual gains in wage over basic demographics

\[
\Delta \log w_i = \beta_2 X_i + \epsilon_i
\]

where \(\Delta \log w_i = \log w_i - \log w_i\) and \(X_i\) same as in Equation 1.19. The coefficients \(\beta_2\) from this regression are the part of the wage gap that would be closed, if commuting and moving costs were equalized at the lowest level, for these workers. The results of this exercise are presented at Figure 1.16. The figure plots the full gap (\(\beta_0\), blue), the portion that can be attributed to the OOI (\(\beta_0 - \beta_1\), red), and the part that will be closed by equalizing commuting costs at the minimal level (\(\beta_2\), yellow).

Differences in commuting costs seem to explain all of the gender gap that is driven by differences in options. Equalizing commuting costs would increase wages for women by about .07 log units, relative to men. This is one third of the overall gender gap. Even though we find that men and women sort into different jobs, there seems to be a similar number of jobs for males and females, therefore the only difference is the distance in
Figure 1.16.: Effect of Commuting/Moving Costs

Note: Blue bars (Total gap) are derived from the coefficient on the corresponding category in a regression of log wages on Male, Citizen, indicator for secondary-education category, a quadratic in age, district density and an indicator for part-time job. The red bars (Total gap from OOI) is the difference in coefficient between the same regression and one that control for the OOI with a coefficient fixed to .26, which was estimated with the two quasi-random variations. The yellow bars (Gap from Commute) is calculate from a similar regression, replacing the dependent variable with minus the gains from reducing commuting costs to their minimal level (see Section 1.7.2 for more details). The reference workers is a female, non-citizen, with intermediate secondary education and 18 years old.
which workers are searching for jobs. This does not mean that there aren’t other ways to increase the OOI for women, such as increasing the supply of jobs that women typically sort into.

In contrast, equalizing commuting costs increases the wage gap between German citizens and non-citizens. These results are surprising at first glance because we found that non-citizens are more sensitive to distance. However, they can be explained by the fact that non-citizens are more concentrated in large German cities. As a result, their commuting costs are already low. German citizens are more dispersed across rural areas, and are more dependent on their ability to commute to jobs in major cities.

The education gap in OOI actually reverses once we equalize commuting costs: workers with intermediate-secondary education have more options than those with higher-secondary. Therefore, the higher-education premium drops by .15 log units (51% of the overall premium), which is more than the full effect of the OOI difference between these groups (.07 log units). This implies that, in a given area, workers with intermediate-secondary education have more relevant job options than workers with high-secondary education. It is only because higher-secondary education workers are willing to take jobs in more distant areas, that they end up with more options. This result can be explained by the fact that more educated workers tend to be more concentrated in occupations that have more industry specific skills, as shown in Figure 1.7. Additionally, intermediate-secondary workers can take both higher-skill, and lower-skill jobs in addition to staying at the same level. While high-secondary workers usually have fewer options to climb to jobs requiring even more skills.

Other than geographical distance, the most significant factor in determining a worker’s match (Appendix Table 1.15) is their training occupation. Workers tend to stay in occupations similar to the ones in which they were trained. Our results in Section 1.5 show that those who undertook training in occupations with more transferable skills have more op-
tions than those who received more narrow training. Since transferable skills are more common in low-paying occupations, the OOI is reducing inequality between occupations.

1.8. Conclusion

In this paper we provide a distinctive and micro-founded approach to empirically estimate workers’ outside options, and to measure the impact of outside options on the wage distribution. The starting point for our analysis was a two-sided matching model, which produced a sufficient statistic for the impact of outside options on wages, the OOI. We took the OOI to the data to identify the workers with better outside options. We then combined this result with a causal estimate of the elasticity between the OOI and wages, to assess the overall impact of options on wages. Our results suggest that differences in outside options generate lower income for females by six percent, non-citizens by eight percent and intermediate educated workers by seven percent (compared to high educated).

Our results indicate that policies that improve workers’ options, including investments in transportation infrastructure or regulation of working hours, are likely to have significant general equilibrium effects. While such policies are usually analyzed only through their impact on workers that directly benefit from them, our results indicate that these policies will likely have important spillovers onto other workers through their outside options. These general equilibrium channels can be studied through their effect on the OOI.

One interesting direction for future work would be to use this framework to analyze specific industries in which outside options play a key role, and good micro-data is available. A similar analysis could also be done on the employer’s side of the market, analyzing heterogeneity in the availability of options for firms, and the impact of outside options on
profits. Finally, the ability to identify workers with better outside options could be useful in studying heterogeneous effects of various policies, or labor market shocks. Our analysis of the heterogeneous response to mass-layoff is one example for how this can be done.
2. Getting Beneath the Veil of Intergenerational Mobility: Evidence from Three Cities

2.1. Introduction

Concerns about poverty and intergenerational mobility are as old as civilization itself.1 Thousands of years later, poverty and intergenerational mobility are, still, among the most important economic and social issues of our time. In an average OECD country it would take four to five generations for children in the bottom earnings decile to attain the level of mean earnings, but there is significant heterogeneity across countries and between ethnic and racial groups within countries. OECD’s annual report on social mobility estimates it will take two generations for children in the bottom decile in Nordic countries to reach the mean; four to six generations across Europe, and five generations in America (OECD, 2018a).

Within the US, differences across racial groups are stunning. For some, America is the

---

1In the oldest written text, “Gilgamesch”, there are mentions of famine and descriptions of poverty are in Confucius writings, the Iliad, and the Odyssey. Inequality and mobility were discussed in Ancient Egypt during the reign of King Akhenaten – 80% of the wealth of the belonged to 20% of the population.
land of unparalleled opportunity. For others, it is the land of the ineludible poverty trap. In parallel work, Chetty et al. (2018) employ a large new dataset – linking census data covering the U.S population to federal income tax returns from 1989 to 2015, a total of 20 million observations – to estimate intergenerational mobility in America, by race and for granular geographies, providing more precise estimates than Solon (1992).² They find that black Americans have lower rates of upward mobility and higher rates of downward mobility compared to white Americans and on par with Native Americans. Blacks in the top income percentile have male children who are as likely to be incarcerated as whites in the 20th percentile.

Gaining a better understanding of intergenerational mobility is of great importance. If, by accident of birth, certain individuals are not able to achieve their full potential then there are important imperfections in the market for talent and making that market more meritocratic may have large social value.

A wide variety of possible correlates of intergenerational mobility have been put forth. These include education and school quality (Aldaz-Carroll and Morán, 2001; McKernan and Ratcliffe, 2005; Isaacs et al., 2008; Baum et al., 2013), neighborhood quality (Keels et al., 2005; Sanbonmatsu et al., 2012; Chetty et al., 2014, 2016; Chetty and Hendren, 2018), early childhood and adverse childhood experiences (Duncan and Rodgers, 1988; Aldaz-Carroll and Morán, 2001; Metzler et al., 2017), family structure and parenting (Duncan and Rodgers, 1988; Aldaz-Carroll and Morán, 2001), and church going and other forms of social capital (Freeman, 1986; Sharkey and Torrats-Espinosa, 2017; Western and Pettit, 2010).

In this paper, we attempt to shed new light on the correlates of mobility in America using new data from in-person interviews of approximately one thousand families – all of

²Solon (1992) used the PSID dataset and corrected for measurement error to show that the correlation between fathers’ earnings and sons’ earnings is approximately 0.4. This proved that US was not as highly mobile a society as was previously believed – previous estimates of the correlation were of the order of 0.2.
which self-report growing up poor – in Memphis, Tennessee, New Orleans, Louisiana, and Tulsa, Oklahoma.

Interviews were conducted in a respondent’s home or a public place, whichever was preferred, and lasted almost two hours. At the end of the interview, the respondent received a pre-paid $150 Visa gift card. Our final sample consists of 928 respondents for the full interview.

The extensive length of our face-to-face interviews allowed us to collect a wide-ranging set of data: basic demographics, mental health, physical health, parental behavior, home environment, childhood experiences, risky attitudes as a teenager, lifetime traumas, neighborhood safety, and psychological skills such as the Big 5 personality traits, grit, locus of control and resilience. Our main outcome variable is log household income in 2016, though we also present results for individual income, adult mental and physical health, and drug and alcohol use.

The results we obtain from these new data are informative, surprising, and inspire the development of new methods. Using typical descriptive approaches, such as those implemented in Garces et al. (2002), the correlates of intergenerational mobility are education, resilience, mental health before age 16, trouble with police before age 18 and grit. Variables such as childhood abuse, parenting, teenage risky behaviors, trust in adults, or other psychological skills are seemingly not important. Moreover using standard descriptive methods to identify which variables are important, when the number of covariates is high, tend to generate results that are not robust to minimal changes in sample size or variable definitions (Mullainathan and Spiess, 2017).

To screen potential survey respondents, individuals were first contacted by phone. A respondent was deemed eligible if: (1) they resided in a zipcode in our desired geographies; (2) were at least 18 years of age; (3) self-identified as having “grown up poor” (e.g. answered “Do you consider yourself to have grown up poor?” in the affirmative). This initial conversation lasted, on average, six minutes, and allowed us to collect information on basic demographics, education, and preferred contact information for all eligible respondents. In total, 457,317 phone calls were made and 6,459 were completed.
This inspired us to think about the approaches used in the literature. Least Squares estimates or popular supervised learning algorithms both assume one can alter individual characteristics in any way the data suggest is optimal – even if one rarely, if ever, observes individuals in the data with those combinations of characteristics. Consider the following thought experiment. Imagine that graduating from college has large causal effects on future income for everyone, but it’s extremely rare that individuals who have endured childhood abuse graduate from college. Just estimating typical descriptive methods – trying to predict income conditional on observables – may tell us to simply get everyone to graduate from college and this will improve their income. But it won’t tell us if it’s possible!

We relax this assumption and use the distribution of covariates to understand how our variables relate to one another and estimate the costs of altering any combination of individual characteristics. In the thought experiment above, this implies that if we observe the lack of childhood abuse and college graduation typically covary, we infer that it is difficult to ensure that victims of childhood abuse graduate from college. And, thus, any income-increasing intervention may want to target both. Our approach boils down to a simple maximization problem with two key parameters: the distribution of individual characteristics and the distribution of the outcome variable, conditional upon those characteristics. We provide both parametric and non-parametric estimates of the correlates of income mobility using our newly collected data and new methods. We believe this method can be used anytime one wants to use observational data to better optimize social experiments designed to increase some desired outcome (e.g. test scores, labor market participation, income in developing countries).

When we use the distribution of characteristics to inform how variables move together, the correlates of intergenerational mobility are quite different. Education is still the most important factor in intergenerational mobility. Of the eight other correlates that are sig-
significant, however, five of them are psychological skills: resilience, the Big 5, self-esteem, self-control, and grit. The remaining three are whether the respondent was in trouble with the police in their youth, the number of adverse childhood experiences – such as abuse – and the number of adult relationships they trusted during their childhood. The fact that education is the most persistent correlate of mobility is consistent with more than a half century of scholarship in economics. Beginning with Blau and Duncan (1967) and formalized in Becker and Tomes (1979), the model of mobility has been the quantity of skills and their prevailing market price. Our results, along with the burgeoning literature on the importance of non-cognitive skills, suggest a much larger set of skills and experiences are important to produce income. Admittedly, the importance of psychological skills in the production of earnings is not a new idea – Heckman et al. (2006) state that “...for many labor market outcomes, a change in non-cognitive skills...has an effect on behavior comparable to or greater than a corresponding change in cognitive skills” – but its relative importance to human capital and family environment is striking.

Variables that capture place-based heterogeneity such as the neighborhood specific probability of the bottom 25 in top 20 percentile are not significant in our main specifications. However, consistent with Chetty et al. (2018) we find that high mobility zipcodes are significantly more helpful to whites. While are able to replicate the positive correlation of such variables with adult income, all of our specifications suggest that these variables are not among the important factors in intergenerational mobility. A potential explanation of this result is that its possible to increase mobility by moving individuals to geographies with more fathers present or higher average mobility, but it may be more effective to alter their human and psychological and human capital directly.

We explore the robustness of our results across various measures of income, alternative specifications, and alternative measures of adult well-being. Our results are virtually unchanged when using adjusted household income or individual income as outcomes
variables. Alternative specifications, such as allowing for heterogeneity in the distribution of individual characteristics or not adjusting for poverty levels as a child, also yield similar results. Interesting differences do emerge when we use mental illness, or drug and alcohol use as alternative measures of adult well-being. The most important correlates of mental illness include various psychological skills, risky behaviors as a teenager, mental and physical illness as a teenager, education, adverse childhood experiences, family environment, interactions with police, number of adult relationships trusted, neighborhood mobility and an index of neighborhood safety. The latter is consistent with data gleaned from the Moving to Opportunity experiment (Sanbonmatsu et al., 2012).

Similarly, the most important correlates of adult drug and alcohol use are self-control, risky attitudes as a teenager, whether an individual was in trouble with police as a youth, mental health, relationship with parents, neighborhood safety, family environment, and adverse childhood experiences. While education was the most persistent predictor of income, it is a significantly less important predictor of adult mental illness or drug and alcohol use as an adult.

Our analysis has three important caveats. First, the locations are not representative. We chose them because they are representative of a significant share of black poverty in America. Appendix Table B.4 compares our survey sample to a comparative sample of National Longitudinal Survey of Youth 1979 (hereafter NLSY) respondents who were between the ages of 14 and 22 and below the poverty line in 1979. Strikingly, our sample – taken from these three cities – is statistically similar to the unadjusted nationally representative data on most demographics. The only variables that are statistically different are age and percent Hispanic.

Second, our survey design requires adults to remember their childhood details with a degree of specificity and objectivity that may be implausible.\(^4\) Adding to this compli-

\(^4\)There is some evidence that individuals are more likely to discuss sensitive topics such as abuse with
cation, our variables designed to assess psychological capital and beliefs – grit, growth mindset, resilience, and so on – are contemporaneous. It is plausible that increasing one's income from childhood poverty causes one to remember childhood experiences in a different light, tell oneself a different narrative or feel more resilient. We had to choose between our current design, which has important shortcomings, and starting a longitudinal dataset similar to the NLSY and waiting at least two decades for the results.5

One way to try and understand how this may affect our results is to analyze any individual characteristics in our data that may have been sampled multiple times in a known longitudinal dataset. NLSY, for instance, assessed the locus of control of its respondents in 1979 when individuals were 14 to 22 and again in 2014 when they were older adults. The correlation between locus of control and income in our data is equivalent to correlating 2014 income with 2014 measures of locus of control in the NLSY; both are positive and significant. Importantly, this correlation in the NLSY exists and is of similar magnitude whether one correlates the 2014 measure of locus of control with 2014 income or the 1979 measure of locus of control with 2014 income. Put differently, it seems that there is a strong correlation between measures of locus of control in youth and adult income and that correlation is almost identical if one uses a measure of locus of control assessed in adulthood.

Third, which is less caveat than clarification, our results are correlates of income and other measures of adult well-being which suggest what types of interventions will be most successful – not causal estimates.6 Without a well-powered field experiment or valid

---

5Another possibility was to ask parents or friends about the respondent when they were young. This might provide a measure of psychological capital when the respondent was young or a different perspective on sensitive issues such as abuse. This method is similarly problematic and, for a fixed budget, reduces our sample size significantly. We chose a larger sample.

6Yet, evidence from experiments or natural experiments, suggests that 7 of our identified correlates of income mobility seem to have a causal effect as well. For instance, Angrist and Krueger (1991) demonstrate that compulsory schooling laws increase years of schooling and hence, higher wages. Heckman et al. (2013) show that participants of the HighScope Perry Preschool Program have higher measures of the Big 5 per-
instrument, thorny issues of self-selection, omitted variable bias, and the like may influence our results. Yet, a separate set of issues exist with field experiments. They produce clear causal estimates but, without understanding the underlying production function of mobility, may be unlikely to yield positive results. We view our approach and field experiments as strong complements. This paper is an important first step in a larger research agenda whose aim it is to take these correlates and conduct a large scale randomized control trial with a subset of them, aiming to increase intergenerational mobility and overall life outcomes of those who, by circumstance of birth, are more susceptible to continue to be low income in future generations.\(^7\)

With the above caveats and clarifications in mind, our paper makes three contributions. First, we collect new detailed data on individuals who were born into poverty. This data is more comprehensive than previous datasets, including information on sensitive issues such as abuse, relationships with parents and other adults, interactions with police, mental and physical health, learning disabilities, and so on. Second, since Fisher (1925), randomized control trials (RCTs) have grown tremendously in use and importance.\(^8\) The methods developed provide a way to use rich observational data to potentially make those experiments more effective, which could save billions of dollars and alter millions of lives. Third, the results from our new data and new approach, offer an innovative way forward for increasing intergenerational mobility in America.

The paper is organized as follows. The next section details our sample frame and the data collected for our analysis. Section 2.3 describes our methods and section 3.5 reports

---

\(^7\) A similar approach was used in Dobbie and Fryer (2013) and Fryer (2015) in an effort to increase student achievement.

\(^8\) The intellectual development of RCTs is varied. Many are theory driven – testing important social scientific theories in the field (e.g. The impact of teacher specialization on student achievement, Fryer (2018)). Others seem more resource driven – the federal government spends $565 billion per year on medicaid and we don’t know how effective these investments are Finkelstein et al. (2012). And, many are impact driven – understanding how best to increase student achievement, employment rates, income in third world countries, or reduce crime. Our method applies to the last category.
empirical results from combining the new methods and data. Section 3.7 discusses our findings in the context of the previous literature and concludes. There are 6 appendices. Appendix B.1 and B.2 are guides to the implementation of our pilot survey and full survey. Appendix B.3 has data descriptions of all datasets used in the paper and their variables. Appendix B.4 contains all proofs. Appendix B.5 and B.6 show the phone and paper survey questionnaires.

2.2. A New Survey of Intergenerational Mobility in America

2.2.1. Design

In choosing which cities to conduct our survey, we were interested in selecting areas with high levels of poverty, ethnic diversity, and geographic variety. We settled on the general areas of Memphis, TN, Tulsa, OK, and New Orleans, LA. To more precisely define the areas, we started with the Metropolitan Statistical Areas that contained these cities and then selected four counties within them: Shelby County, Tulsa County, and Jefferson and Orleans Parishes. Shelby County has a population of 936,961 with 54.1% black, 6.4% Hispanic, and 35.9% non-Hispanic white. Twenty-one percent of the population currently lives in poverty. Tulsa County has a population of 646,246 with 10.8% black, 12.7% Hispanic, and 60.2% non-Hispanic white. Sixteen percent of the population currently lives in poverty. Since the counties in the New Orleans MSA had smaller populations, we selected two: Jefferson and Orleans Parishes. Jefferson Parish has a population of 439,036 with 27.6% black, 14.9% Hispanic, and 52.5% non-Hispanic white. Sixteen percent of the population lives in poverty. Orleans Parish has a population of 393,292 with 60.1% black, 5.7% Hispanic, and 30.7% non-Hispanic white. Twenty-four percent live in poverty.
2.2.2. Sample Selection Method

After determining our sample areas, the next step was to decide how to select our sample within those areas. Previous surveys have often relied on address-based sampling (ABS) to help ensure a representative sample. For example, in the National Longitudinal Survey of Youth 1979, interviewers went to a random sampling of housing units and performed a short screener in person. Although this method is often considered the gold standard for in-person interviews, it is also very expensive. An alternative method of screening is by phone. In order to determine if the samples obtained by phone-based screens and ABS were comparable, we ran a pilot study in Los Angeles County, California, consisting of 643 residents. After comparing the two samples and finding no significant differences in the demographics of the populations surveyed, we selected the phone screening method as it is much more cost-effective and allowed us to interview more individuals for our final sample. (See Appendix B.2 for a full methods report on our survey design). Abt Associates was responsible for the implementation of both the pilot and full interview.

2.2.3. Phone Screens

In order to be eligible for the full interview, individuals had to reside in a zip code that was in one of our sampling counties, be at least 18 years old, and self-identify as having grown up poor. Sixty-five percent of the screening frame came from a cell phone screened sample, and thirty-five percent came from a landline screened sample. During the phone screen, we collected information on basic demographics including gender, race, education and contact information for those individuals who were eligible. (See Online Appendix B.5 for the full text of phone screen). The phone screens lasted an average of 6.3 minutes. In total 6,459 phone screens were completed in our three sampling areas: 1,227 were eligible and agreed to participate in the full interview, 1,390 were eligible but refused fur-
ther participation and 3,842 were ineligible. (See Appendix B.2.5 for full distribution of respondents by location).

2.2.4. Interviews

For subjects who were eligible and agreed to participate, in-person interviews were scheduled. These interviews lasted an average of 104 minutes; approximately 350 questions were asked. The majority of interviews were conducted in individuals’ homes, although interviewers were also willing to meet with respondents in public places like coffee shops or the library if the respondent preferred. At the end of the interview, the respondent received a prepaid $150 Visa gift card. During the interviews, we asked questions on a wide variety of topics. We incorporated topics that have been considered potential causes of poverty in the literature and focused on questions and psychological scales that have been developed and validated in previous studies. Below, we describe the general categories of topics we were interested in along with some of the major subsections.  

2.2.5. Income and Adult Well-Being

The main outcome variables used in our analysis are (log) individual income and household income in 2016 in dollars, a series of detailed questions designed to assess mental health, and a set of questions to assess drug and alcohol abuse. Of 928 respondents, 764 were willing and able to answer the income question in an open-ended format. If a respondent said that she did not know or want to answer, we asked her if she would be willing to tell us what range it fell within ($0-$10,000; $10,000-$20,000; $20,000-$30,000; $30,000-$40,000; $40,000-$50,000; $50,000-$75,000; $75,000-$100,000; $100,000-$150,000 or more than $150,000) and then assigned her the midpoint of that range.

The full interview text is available in the Appendix B.6.
Beyond income, we analyze two additional measures of adult well-being: mental illness and drug and alcohol abuse. To assess these measures, we used the industry standard instruments for screening in clinical settings. This includes CAGE-AID: a questionnaire developed to screen for drug use (Basu et al., 2016); GAD 7: a 7-item self-administered questionnaire developed to measure and assess generalized anxiety disorder (Spitzer et al., 2006); and PHQ-9, a 9-question instrument used for screening, diagnosing, monitoring, and measuring the severity of depression (Spitzer et al., 2000). These instruments have been used widely in research and clinical practice and generally have been shown to have superior validity when compared to alternative screening questionnaires as well as reliability with independent diagnoses conducted by mental health professionals (Kroenke et al., 2001; Löwe et al., 2004; Brown et al., 1998; Brown and Rounds, 1995; Spitzer et al., 2006). See Appendix B.3 for details.

2.2.6. Other Variables

Basic Demographics

We collected a number of demographic variables in both the phone screen and full interview. This includes gender, race (coded as white, black, Hispanic or other), age, current household members, and highest level of education completed. Our education variable asked individuals to classify themselves as not completing high school, having a high school degree/GED/some college, having a two year Associate’s degree, or having at least a Bachelor’s degree.

Parental Income

Although we do not contain actual estimates of parental income, we collected variables that proxy for parents’ economic conditions. These include participants’ responses to questions about whether their families were well off, average or poor financially; whether
they ever moved due to financial difficulties or ever received help due to financial difficulties; if there was a time when the father was unemployed for several months; if the mother ever received welfare; and the frequency with which their families (a) found it difficult to afford child care, (b) fell behind rent or mortgage payments, (c) fell behind gas, electric or phone bill payments, (d) were unable to pay for transportation to get to work or school, (e) were unable to afford medical care, (f) had trouble paying a credit card balance, and (g) had too little money to buy enough food. To ensure that we are targeting people who grew up poor only, we drop participants who say that they had grown up “well off”. This reduces the sample size by 28 respondents.

*Early Home Environment*

These variables were meant to capture the environment that an individual was raised in through questions about childhood household composition, neighborhood safety, and financial difficulties in their childhood. Questions were mainly drawn from surveys administered as part of the Moving to Opportunity experiment and the Health and Retirement Survey. We were particularly interested in looking at the roles and practices of parents or parental figures and other important relationships in childhood. We relied on the Short Version of the Family Environment Survey, a 27-item inventory designed to measure social-environmental characteristics of the family that was developed by Rudolf Moos and Bernice Moos in 1994. This scale features items such as “Family members had strict ideas about what is right or wrong” and “There were very few rules to follow in our family” and asked respondents to state whether these statements were true or false of the family they lived with between the ages of 5 and 12. Additionally, we adapted questions from the Parent Practices Survey, a 34-item self-reported instrument designed by Dr. Joseph Strayhorn to understand parents’ patterns of interaction with their preschool children. In a sample of 200 low-income parents, the scale had good internal consistency and 6-month stability (Strayhorn and Weidman, 1988) and was associated with measures of
parents’ psychological and social health. Since the original survey was targeted towards parents, we reworded questions to ask respondents their perceptions of their parents’ attitudes. For example, “How often does this child do something that gives you pleasure and enjoyment” was rewritten as “How often would your parent say that you did something that gave him/her pleasure and enjoyment?”

**Childhood Traumas and Risky Behaviors**

Our main instrument for assessing childhood trauma is the questionnaire from the Adverse Childhood Experience (ACE) study, a 10-item self-reported measure developed to quantify cumulative childhood stress (Felitti et al., 1998). The study, and further follow ups, repeatedly found a relationship between negative later life health outcomes (e.g. alcoholism) and increasing numbers of stressors (e.g. an ACE). ACEs include items such as whether an individual experienced childhood physical or sexual abuse, negligence, witnessed the physical abuse of another household member or whether another household member was mentally ill or addicted to drugs. We investigate whether this relationship continues to be robust in a population of mainly low income individuals and if applied to outcomes outside of the health domain. Additionally, we used a list of lifetime traumas that an individual may have experienced before the age of 18 (such as being in trouble with the police, repeating a year of school or being abandoned by his parents) along with questions on childhood health that were taken from the Psychosocial and Lifestyle Questionnaire of the Health and Retirement Survey. Questions about an individual’s experience with drugs and alcohol during childhood were drawn from the 2013 National Youth Risk Behavior Survey. Finally, questions about risky attitudes and behaviors during adolescence were taken from the Moving To Opportunity Child Questionnaire.

**Psychological Skills**

We chose a broad spectrum of psychological scales, some of which have been tested in
large surveys previously (e.g. Rotter’s locus of control scale), as well as some that have been developed more recently (e.g. Grit Scale, Dweck Mindset Instrument). Where available, we used abridged scales to limit the length of the survey. Below, we describe a selection of these scales.

The Brief Resilience Scale is a 6-item scale developed by Smith et al. (2008) to assess the ability to recover from stress. It has been examined in samples of students as well as chronic pain and cardiac patients and found to be reliable and negatively related to mental and physical health symptoms including anxiety and depression. The scale includes items such as “I tend to bounce back quickly after hard times” and asks respondents to state whether they strongly disagree, disagree, neither agree nor disagree, agree, or strongly agree. A meta-analysis of various resilience scales used in the field found that the Brief Resilience Scale was among one of the highest rated scales in terms of construct validity and internal consistency (Windle et al., 2011).

The Brief Self-Control Scale (BSCS) was developed by Tangney et al. (2004) and is used to measure the five domains of self-control: self-discipline, resistance to impulsivity, healthy habits, work ethic and reliability. The scale was found to have good internal consistency and retest reliability. Higher scores on the self-control scale were correlated with lower rates of alcohol abuse, higher grade point averages and better interpersonal skills. The BSCS has been used in over 100 published studies on adolescents and adults to predict numerous behavioral outcomes including high school achievement, job-searching behavior, binge eating and work performance (Lindner et al., 2015; Duckworth and Seligman, 2005; Baay et al., 2014; de Ridder et al., 2012).

Rotter’s locus of control scale is a 23 item scale developed by Julian Rotter in 1966 to assess the extent to which an individual feels he can control his circumstances and outcomes. We used an abridged version that relied on four items to mirror the version used in NLSY79. Each item contains two statements (e.g. A. In my case getting what I want has little or
nothing to do with luck and B. Many times we might just as well decide what to do by flipping a coin.) and asks respondents to select which statement is closer to their opinion. The four-item version used in NLSY was found to correlate with schooling decisions, employment and wages (Heckman et al., 2006; Goldsmith et al., 1997).

The Dweck Mindset Scale was developed by psychologist Carol Dweck and is used to differentiate between a fixed mindset, in which individuals believe basic qualities like intelligence are fixed traits, and a growth mindset, in which individual believe that their abilities can be developed. A study of high schoolers in Chile found that having a growth mindset strongly predicted academic achievement, particularly among low-income students (Claro et al., 2016). Similarly, a longitudinal study of middle school students found that those with a growth mindset outperformed students with a fixed mindset in mathematics two years later (Blackwell et al., 2007). We used an abridged 3-item version that focuses on fixed views of intelligence such as “Your intelligence is something about you that you can’t change very much” (Dweck, 1999).

The Rosenberg Self Esteem Index is a 10-item scale developed by Dr. Morris Rosenberg in 1965 to measure both positive and negative feelings a respondent may have about himself. It has been widely used across fields and in large surveys including in the NLSY79. In the NLSY79, higher measures of self-esteem were correlated with future economic success (Heckman et al., 2006). Respondents are asked to indicate their level of agreement with statements such as “On the whole, I am satisfied with myself" and “I feel I do not have much to be proud of."

The original 12-item Grit Scale was developed by Duckworth et al. (2007) and is meant to measure the ability to sustain effort towards long-term goals. We used the 8-item Short Grit Scale which has been found to correlate with retention in cadets attending West Point, higher educational attainment and fewer job changes among adults (Duckworth et al., 2007; Duckworth and Quinn, 2009). The 8-item scale contains statements like “New ideas
and projects sometimes distract me from previous ones” and “I often set a goal but later choose to pursue a different one” and asks respondents to indicate their agreement with the statement on a five-level Likert scale.

To assess personality traits, we first started with the the International Personality Item Pool, a site with over 3,000 items and 250 scales that are used to measure personality traits (Goldberg et al., 2006). We selected the 50-item sample questionnaire based on Goldberg’s markers for the Big-Five domains of personality: extroversion, agreeableness, conscientiousness, emotional stability, and intellect (Goldberg, 1992). The 50-item Big-Five scale (IPIP) has good internal consistency and related strongly to two other leading personality questionnaires – NEO Five Factor Inventory and Eysenck Personality Questionnaire Short Form (Gow et al., 2005). Respondents are asked to indicate how accurate they think statements such as “I am interested in people” or “I pay attention to details” are in describing themselves.

### 2.2.7. Descriptive Statistics

Our final sample consists of 900 respondents. Appendix B.2.5 provides brief sample accounting. We made 458,317 phone calls; 6,459 completed the initial phone screen; 2,617 were eligible for the full study and 3,482 were deemed ineligible. 1,227 agreed to be surveyed and 75.63% of them actually completed the survey.  

Appendix Table B.5 compares our sample to participants of the 2016 American Commu-
nities Survey (ACS), a sample of 3,156,487 individuals from across America. The mean household income for our set of respondents is $47,484, compared with $91,850 nationally. We have a significantly higher fraction of black people – 43.7% in our sample compared to 12.7% in the national average. Conversely, we have a lower fraction of hispanic people – 9.7% in our sample compared to 17.8% in the national average. Our sample is statistically similar to the national average in terms of gender. As one might expect, our sample is less educated than the national average – 48.8% have incomplete college degrees compared to 40.7%, while only 20.8% of respondents have a bachelor’s degree or higher in our sample compared to 23.1% nationally. In other words, our sample is more likely to be poor, has a higher fraction of black people and is less educated.

We also compare our sample to a sample of individuals from the National Longitudinal Survey of Youth (NLSY) who experienced poverty in their youth. In contrast to the ACS sample, our sample and the subset of those in the NLSY who experienced poverty in their youth look quite similar.11 Appendix Table B.4 compares 4 demographic variables including individual income as well as some questions that assess locus of control and mental illness that are both contained in our sample and the NLSY. Column (1) contains summary statistics from our sample. Column (2) presents these statistics for the sample of the NLSY who experienced poverty in their youth. Column (3) is identical to column (2) but reweights the observations so that the NLSY data has the same distribution on 3 exogenous variables – age, race, and gender – as our sample.

We begin by comparing our sample to the unadjusted NLSY sample. The average age in our sample is 49, and 53 in the NLSY sample. The difference, 4 years, is statistically significant. As NLSY gathers information on individual and family income only, we compare individual income across both samples. The mean individual income in our sample is

11We say an individual “experienced poverty in their youth” in the NLSY sample if a respondent from 2014 was between the ages of 14 and 22 in 1979 and the survey reported that their family was below the poverty line in 1979.
$28,312 compared to $27,751 in the NLSY sample. The difference, $561, is not statistically significant. Other demographic variables such as gender are also statistically similar. Our sample has more blacks and less Hispanics than the NLSY. We also compare questions from the adult mental illness index that overlap between the two surveys. Our sample has worse adult mental illness on three of the four subcategories and lower self-esteem scores than the NLSY.

Comparing our sample to the adjusted NLSY sample gives similar results except on individual income and fraction of black people. Mean individual income in the adjusted NLSY sample is $24,247 and is statistically smaller than the mean in our sample. Mean age in the adjusted NLSY sample is 53 and statistically larger than the age in our sample, while the percentage of women is 52 and statistically similar. While the adjusted NLSY sample has more Hispanics, it has statistically similar percentage of blacks compared to our sample. With regard to adult mental illness and other psychological traits, the adjusted NLSY sample looks similar to the unadjusted NLSY sample i.e. the sample has better adult mental illness on three of the four subcategories and higher self-esteem scores compared to our sample.

2.2.8. Correlations with Income

To get a sense of how these new data – unprocessed – correlate with income, Appendix Figures B.1-B.5 displays binned scatter plots of household income on a set of 18 indices individually, which together, encapsulate all 350 questions asked of our respondents. All figures plot a scatter graph, as well as a quadratic fitted line, of the log of household income on binned categories of indices with each observation weighted by its associated survey weight.

The general picture that emerges is that adult income is strongly correlated with mental
illness before 16 years of age, psychological traits, family environment, lifetime traumas and neighborhood mobility.  

Appendix Table B.7 displays pairwise correlation coefficients between any two given indices and provides a good sense of how potential covariates of intergenerational mobility might relate to each other. Of the several coefficients presented, the most noteworthy ones are between adverse childhood experience and mental illnesses before 16 years of age (0.426), adverse childhood experience and traumas experienced before 18 years of age (0.517) and adverse childhood experience and parenting (-0.572). Good parenting is highly correlated with better relationship with parents (0.510) and, consistent with Chetty et al. (2018), neighborhood mobility is highly correlated with fraction of fathers present.

---

12 Appendix Figure B.1 displays the relationship between income and four health-related indices: (a) mental illness before 16, (b) physical illness before 16, (c) psychological index, and (d) diet. Mental and physical illness indices were created as totals of standardized responses to questions on “Childhood Health Experience”. Psychological index is the total of 7 sub-indices: resilience, locus of control, growth mindset, grit, self esteem, Big 5 personality traits, and self control. Diet is the total of standardized responses to 2 questions – fraction of days the respondent received three meals and fraction of days he received a balanced diet between the ages of 5 and 12. A detailed description of how each index was created is given in the Appendix B.3.

Both mental illness before 16 and the psychological index show strong correlations with income – income declines as mental illnesses in childhood increase and income rises steeply with more psychological skills. Appendix Figure B.2 similarly plots six indices which, together, provide a reasonably comprehensive picture of each respondent’s childhood family environment. Income is positively correlated with the quantity of adult relationships trusted in their childhood, and is also positively correlated with the quality of those relationships – individuals who could trust their parents had higher income than those who did not trust their parents but trusted a teacher or coach instead. Family environment – which encompasses true/false statements such as “we fought a lot in our family”, “family members were rarely ordered around”, “we didn’t say prayers in our family”, “we didn’t believe in heaven or hell”, “family members sometimes hit each other”, “everyone had an equal say in family decisions” etc. – is positively related to income and so too, is the quality of a family’s social network (defined as the total of standardized responses to questions on if parents’ close friends lived in the same neighborhood, graduated from college, worked full time or were a different race). Surprisingly, parenting skills and relationship with parents was not directly correlated with income. These variables are strongly correlated to family environment, however.

Appendix Figure B.3 displays the same graphs for five childhood experience indices – (a) adverse childhood experience (ACE), (b) risky attitudes as a teenager, (c) trauma before 18 years of age, (d) any trauma in lifetime, and (e) beliefs about success in life. Income levels decline when teenagers exhibit more risky behaviors or when individuals report more lifetime trauma. Surprisingly, however, the relationship between income and adverse childhood experiences is not significant, though ACE and youth mental health have a correlation of almost 0.5.

The final set of graphs are shown in Appendix Figure B.4 which plot income and three neighborhood indices – (a) neighborhood income mobility, (b) fraction of fathers present in neighborhood, and (c) neighborhood safety index. All neighborhood indices are positively correlated with income levels.
Traditionally, social scientists interested in correlates of intergenerational mobility have estimated models of the following form:

\[ income_{t+1} = \alpha + \beta X + \gamma income_t + \epsilon \]  

(2.1)

*Western and Pettit (2010)* use this approach on the National Longitudinal Survey of Youth 1979 data and find that while two-thirds of non-incarcerated low-income men are upwardly mobile, only one in four out of incarcerated men rises out of the bottom quintile of the earnings distribution. *Sanbonmatsu et al. (2012)* use data from interviews of adults from the “Moving to Opportunity” households and infer that moving into a low-poverty neighborhood during childhood has substantial effects on the physical and mental health of adults. \(^{13}\) Most recently, *Chetty et al. (2018)* demonstrate that the only variables that explain racial differences in mobility across geographies (out of 23 analyzed) is the fraction of black fathers present in the census tract and racial bias in the county (measured as scores on Implicit Association Tests and an index based on the frequency of Google searches for racial epithets). The authors estimate that children who grow up in a tract with 10 percentage points more black fathers present have incomes that are 0.5 percentiles higher on average. Conversely, counties with 1 standard deviation higher level of racial bias against blacks have mean income ranks that are 0.8 percentiles lower. \(^{14}\)

Using this traditional approach on our new data yields some interesting but puzzling results, which are depicted in Figure 2.1. Five variables are significantly correlated with intergenerational income mobility – education, resilience, mental illnesses before age 16,

\(^{13}\)Appendix B.7 gives a review of studies that estimate correlates of intergenerational mobility from various data sources.

\(^{14}\)Although, black-white wage gaps are smaller in “good” neighborhoods with low poverty, high rates of father presence and low rates of implicit bias, fewer than 5% of blacks live in such neighborhoods.
Figure 2.1: OLS Method
trouble with police before age of 18 and grit. The results are consistent with McKer-
nan and Ratcliffe (2005); Isaacs et al. (2008); Baum et al. (2013); Aldaz-Carroll and Morán
(2001); Sharkey and Torrats-Espinosa (2017); Western and Pettit (2010).

Surprisingly however, childhood risky behaviors, physical or sexual abuse or neighbor-
hood mobility do not register as significant once one controls for other variables. Of
course, they may be operating through variables such as trust of adults and mental health,
but if we were designing an intervention based on these data and were not confident
about the income production function, one could conclude that increasing education and
giving them resilience training – and ignoring parenting, abuse, or risky behaviors –
would significantly increase income. Moreover, the results are not robust for different
choices of income variables. Using individual income as our outcome, some variables
such as district mobility and beliefs about success had an opposite effect of what we ex-
pected. This is a known problem is using descriptive methods to assess variable impor-
tance with data sets which have many potential explanatory factors (Mullainathan and
Spiess, 2017).

This inspired us to think beyond the traditional approach.

2.3. Methods

In what follows, we develop a method that one can use to understand which factors from
observational data have the greatest potential to increase a pre-specified outcome. In our
particular case, we view these variables as potential levers to change in a field experiment
designed to increase income mobility and adult well-being. Our method unearth factors
that increase mobility directly, as well as the variables that influence those factors, and so
on, without assuming the causal structure among the set of data we collected.
We begin with a simple example we can solve analytically.

2.3.1. An Example

Let there be two types of children reared in poverty: $\theta_0$ and $\theta_1$. Further assume that $\theta_0$ children have low resilience and belong to abusive households. In contrast, $\theta_1$ children have high resilience and belong to non-abusive households.

The probability of graduating from college is different across types: $P(\text{Graduate}|\theta_0) = .05$ and $P(\text{Graduate}|\theta_1) = .95$. We assume

$$income = \alpha \ast \text{graduation} + \epsilon_0$$

Notice, conditional on college graduation, income is independent of type. In other words, a $\theta_0$ that graduates from college can expect the same income as a $\theta_1$.

Traditional methods that don’t take into account the graduation production function, will estimate $\hat{\alpha}$ correctly and will suggest that we should help kids graduate from college. Resilience and childhood abuse don’t matter, as long as everyone gets to go to college and has the potential to graduate. Namely, estimating

$$income = \hat{\alpha} \ast \text{graduation} + \hat{\beta} \ast \text{Abuse} + \hat{\gamma} \ast \text{resilience} + \epsilon,$$

yields $\hat{\alpha} = \alpha > 0, \hat{\beta} = \hat{\gamma} = 0$.

However, $\theta_0$ children graduating from college is rare, potentially because or and abusive background are affecting graduation rate. In that case, any intervention that lowers the costs for them to enroll (e.g. free test prep or more aggressive guidance counselors) or attempts to make their time on campuses more enjoyable will likely fail because it will
not translate into increased graduation. In contrast, if we take into account the graduation production function, we will choose different interventions.

In what follows, we won’t try to estimate the equivalent of the graduation production function, but rather, assume that on average it’s easier to generate variation that we naturally observe in the data. That is, rather than push \( \theta_0 \) children to graduate from college, we will try to make the \( \theta_0 \) children more similar to \( \theta_1 \) children. Interventions motivated by our approach invest in increasing resilience and counseling abused children.

This approach still may not be effective if \( \theta_0 \) children are different from \( \theta_1 \) children on some unobserved characteristics, or observed characteristics that are just impossible to change (e.g. race or genes). In this case, our approach is equivalent to making the \( \theta_0 \) children who do not graduate from college more similar to the 5% of the \( \theta_0 \) that do. Assume for instance that those 5% are much more likely to have trusting relationships with adults, compared to the other 95% that do not graduate from college. Then we would like to focus mostly on ensuring there are adults in a child’s ecosystem that they can trust, as it seems to be particularly important for college graduation of \( \theta_0 \) children.

### 2.3.2. Using Observational Data to Inform Social Experiments

Imagine that we want to improve some outcome \( Y \) (say, log income) and we have data on many observables, \( X \), from a set of individuals. We want to design an experiment that would generate the largest expected increase in income.

Below, we describe a general method to accomplish this based on two assumptions: (1) there is a causal relationship between the set of characteristics in the data, \( X \), and a pre-specified outcome, \( Y \); and (2) the expected costs of an intervention can be approximated by a measure of statistical distance between the pre-intervention and post-intervention distributions of \( X \).
The first assumption ensures there is a problem worth solving – finding correlates that have no potential causal impact is useless. This assumption is not meant to be a reasonable description of reality, but rather something that will be verified in the field experiment. The second assumption is the key innovation of our method. It accounts for the causal links between different $X$ variables in a simple manner. We formalize this intuition below.

**Assumption 2.1.** There is a causal relationship from $X$ to $Y$ such that $Y = f(X) + \varepsilon$ and $f(X) \perp \varepsilon$. If we assumed $f$ was linear, we could estimate this relationship with ordinary least squares regressions, or, if we want to make fewer assumptions, supervised learning algorithms. These approaches, which are standard in the literature, are problematic for the reasons discussed throughout.

Let there be a process that determines $x \in X$, albeit imperfectly. Suppose there is an intervention $I \in \{0, 1\}$ such that $I = 0$ is status quo and $I = 1$ is an intervention. Let the existing (e.g. status quo) distribution of $X$ be denoted $P(X|I = 0)$, and the post-intervention distribution of $X$ be denoted $P(X|I = 1)$. We assume $I \perp \varepsilon$.

We do not assume that one has any data on interventions so the distribution $P(X|I = 1)$ cannot be directly computed.\(^{15}\) Moreover, we do not assume that we have a model of all the potential causal relationships of $X$, which precludes one from writing down a structural model. And, importantly, we don’t assume that we can alter $P(X|I = 1)$ in any way we want.

In lieu of making these typical assumptions, let $C(I)$ denote the cost of an intervention $I$. We assume that, in expectation, this cost is higher the more we change the existing distribution $P(X|I = 0)$. Formally, we will use Kullback-Liebler divergence as a measure of statistical distance between the two distributions (as in Hansen and Sargent, 2001).

\(^{15}\)Precisely, what data would this require.
**Assumption 2.2.** The expected cost of an intervention is the Kullback-Liebler Divergence

\[ E[C(I)] = D_{KL}(P(X|I=1)||P(X|I=0)), \]

where

\[ D_{KL}(P(X|I=1)||P(X|I=0)) = \int_X P(X|I=1) \log \frac{P(X|I=1)}{P(X|I=0)} dX \]

Cost is minimized at \( D_{KL} = 0 \) (when \( P(X|I=1) = P(X|I=0) \)). For all other choices of \( P(X|I=1) \), \( D_{KL} > 0 \) and is increasing when the distributions are less similar.\(^\text{16}\)

Let the expected effect of an intervention, \( I \), on an outcome \( Y \), be denoted \( E[\Omega(I)] \),

\[ E[\Omega(I)] = E[Y|I=1] - E[Y|I=0] = E[f(X)|I=1] - E[f(x)|I=0] \]

Now, suppose we want to find an intervention \( I \) – a distribution of \( X \) – that maximizes the expected effect on income, given some budget constraint: \( \max_{P(X|I=1)} E[\Omega(I)] \) subject to \( E[C(I)] < B \).

In Lagrangian form: \( L = E[\Omega(I) - \lambda (C(I) - B)] \), where \( \lambda = \frac{\partial E[\Omega(I^*)]}{\partial B} \) is the shadow price of an intervention. We assume decreasing marginal returns \(- \frac{\partial^2 E[\Omega(I^*)]}{\partial B^2} < 0\) – which implies that \( \lambda \) is monotonically decreasing in \( B \).

We can simplify the Lagrangian further with two observations. First, \( E[Y|I=0] \) is the status quo, which is constant for all \( I \). Hence, to maximize \( \Omega(I) \) we only need to maximize \( E[f(X)|I=1] \). Second, \( B \) is a constant. Thus, we have the following result:

**Proposition 2.1.** Under Assumptions 1 & 2, the optimal intervention \( I \), solves the following

\(^\text{16}\text{Formally, in information theory, } D_{KL}(P(X|I=1)||P(X|I=0)) \text{ represents the amount of information gained when learning the data is generated from distribution } P(X|I=1) \text{ instead of } P(X|I=0).
maximization problem

\[
\max_{P(X|I=1)} \ E \left[ f(X) | I = 1 \right] - \lambda D_{KL} (P(X|I = 1) \| P(X|I = 0))
\] (2.2)

for a given value of \( \lambda \).

When \( \lambda \to \infty \) it means we can’t change the distribution of \( X \) much (\( B \to 0 \)). When \( \lambda \to 0 \) it means we can change it easily (\( B \to \infty \)).

The solution to this maximization problem yields the joint-distribution of \( X \) after the most cost-effective intervention in expectations. Some of the \( X \) variables will be change in order to increase \( Y \) directly. Other variables in \( X \) will be changed in order to affect \( Y \) through other \( X \) variables. And some \( X \) variables may also change just as a side-effect of the change in the causal variables. While we cannot distinguish between those different cases without additional knowledge, the solution still focuses our search to a much fewer set of interventions.

While conceptually straightforward, this maximization problem can be difficult to solve in practice. We need to estimate two functions: \( f(X) \) and \( P(X|I = 0) \). Estimating \( f(X) \) is a standard problem for which one can apply an array of statistical methods (“supervised learning”). To estimate \( P(X) \), we need to characterize the distribution of the \( X \)s without any outcome variables; an “unsupervised learning” problem. Finally, we need to solve the maximization problem and find \( P(X|I = 1) \) for those estimated values.

In the next sections, we solve the maximization problem under three different sets of assumptions.\(^{17}\) First, we estimate \( P \) assuming \( X \) has a multivariate normal distribution and estimate \( f \) assuming it is linear. The solution in this case is quite simple. We then use empirical likelihood to estimate both \( P \) and \( f \) non-parametrically which provides a

\(^{17}\)We ignore sample weights in this section for clarity. We prove more general versions of each of the theorems with sample weights in Appendix B.4.
solution to our maximization problem allowing for flexibility in the relationship between the variables and the relationship between those variables and our outcomes. Finally, we allow for heterogeneity in the distribution of $X$ across individuals.

### 2.3.3. $X$ is Normal, $f$ is Linear

Assume $X$ has a multivariate normal distribution and $f$ is linear. With these assumptions, we can prove the following result – which greatly simplifies our problem.

**Proposition 2.2.** Assume $f(X)$ is linear, $X \sim N(\mu_0, \Sigma_0)$ and Assumptions 1&2 hold. Assume also that $X$ is normal after the intervention $(X|I = 1 \sim N(\mu_1, \Sigma_1))$. Then for the optimal choice of $I$

$$\mu_1 = \mu_0 + \rho$$

$$\Sigma_1 = \Sigma_0$$

where

$$\nu = \frac{1}{\lambda} COV(X, Y)$$

This proposition simplifies the characterization of the optimal choice for $I$. First, if $X$ is normal and $f(X)$ is linear, the optimal intervention is an increase of $X$ by a constant vector $\rho$. Second, this constant is the covariance of $X$ with $Y$ divided by lambda. If we standardize $X$ and $Y$ (to make them unit free) then it’s proportional to the correlation coefficient. Thus, raw correlations provide the direction, and if the shadow price is smaller, we proceed further in that direction. And, given our focus on the relative importance of each variable compared to other variables, the value of $\lambda$ doesn’t effect the results.

This is a unique case, where we can analytically solve for the optimal intervention without actually estimating $f$ and $P$ directly.
The intuition described in our example shines through: this method may choose to increase some variables even if they do not effect the outcome directly, once controlling for other variables ($\beta_j = 0$).

Recall, $Y = \alpha \text{Graduation} + \varepsilon$ and, now assume a specific linear production function for graduation

$$\text{Graduation} = \beta \text{Resilience} + \gamma \text{Abuse} + \upsilon$$

where $\beta > 0 > \gamma$. Assume that Resilience, Abuse and $\upsilon$ are normally distributed so the assumptions of Proposition 2 hold.

In this world, resilience and childhood abuse would be correlated with income $Y$, but not once one controls for graduation. Designing an experiment to intervene only in college graduation, keeping other things fixed, is an intervention in $\upsilon$. From Proposition 2, our method will describe the most important correlates of the intervention in the X-space, based on their covariance with $Y$. The strongest intervention would likely need to be in graduation as $\text{COV} (Y, \text{Graduation}) = \alpha \text{V(Graduation)}$ which is just $\alpha$ if we normalize all $X$ variables to have standard deviations of 1.

Then, the covariance of resilience can be written as

$$\text{COV} (Y, \text{resilience}) = \alpha \beta \text{V(Resilience)} + \alpha \gamma \text{COV} (\text{Resilience, Abuse})$$

and similarly childhood abuse. We get that the optimal intervention would affect resilience more when $\beta$ is higher, so when resilience has a stronger effect on graduation.

It will also be higher when $\gamma$ and $\text{COV} (\text{Resilience, Abuse})$ are higher in absolute terms: when childhood abuse has a large effect on graduation and they are also (negatively) correlated with resilience. This could be because resilience is decreasing with abuse, which increases graduation even further. Or, because reducing abuse increases resilience as well.
Or, because of another factor, that is increasing both simultaneously. In either case, resilience will increase more in the optimal intervention, either directly or through other channels.

One can also infer how much the optimal intervention would be in the direction of policies that increase graduation directly, while holding other variables fixed, such as free test prep or more aggressive guidance counselors. This is captured in the covariance of $Y$ with the orthogonal part of $Graduation$, which is $v$. In this example, this can be written as: 

$$COV(Y, v) = \alpha V(v)$$

Therefore, our method would show that the optimal intervention aims to increase graduation more when $V(v)$ is higher. This would be the case when there is a lot of variation in $Graduation$ which is not driven by resilience or abuse. That is, if there is a significant variation in graduation rate for people with the same level of resilience and childhood abuse, this policy is more likely to be preferred. The intuition is straightforward - if there are many people that are able to graduate from college with low resilience and abused households, then it seems plausible that trying to improve graduation directly is optimal. In contrast, if it’s very rare to observe college graduates with low resilience and abused childhoods, it implies that it’s very hard to graduate without improving those issues as well. In this case we would want to increase resilience and counsel abused children in order to improve graduation.

### 2.3.4. Non-Parametric Estimation of $P(X)$ and $f(X)$

The assumptions that $X$ has a normal distribution and $f$ is linear are quite strong. We now outline an approach to estimating both $P(X)$ and $f(X)$ non-parametrically, using empirical likelihood (Owen, 2001).

To estimate $P$ we assume that every $x_i$ we observe in the data has a probability of $\frac{1}{N}$,
where $N$ is the sample size. This is the distribution one assumes on the data when using bootstrap methods. The probability to observe an $x$ that is not in the data is set to 0. In symbols:

$$P(X = x|I = 0) = \begin{cases} \frac{1}{N} & \exists x_i \in \text{data s.t. } x_i = x \\ 0 & \not\exists x_i \in \text{data s.t. } x_i = x \end{cases}$$

Thus, an intervention alters the probabilities of the observed $X$ vectors. The KL divergence for any choice of $I$ is $D_{KL}(P(X|I = 1)||P(X|I = 0)) = \sum P(X = x|I = 1) \log \frac{P(X = x|I = 1)}{P(X = x|I = 0)}$.

Our problem simplifies to choosing values for $p_i = P(X_i|I = 1)$ where $x_i$ is the $i$th observation. The solution is attained from the following proposition:

**Proposition 2.3.** Assume $P, f$ are distributed as above and Assumptions 1 & 2 hold. Then for the optimal choice of $I$

$$P(X = x_i|I = 1) \propto w_i^{\lambda - 1}$$

Therefore, the solution is a reweighted distribution, that puts larger weights on higher-income individuals. Notice: higher $\lambda$ means that we can’t change much and probabilities remain similar to uniform. Lowering $\lambda$ puts more weight on high-earners. We use a value of $\lambda = 100$, though similar values yield similar results. In the Appendix B.4 we extend this to a more general case, where the optimal intervention cannot change some characteristics (e.g. race, gender), and show the solution is similar.

### 2.3.4.1. Heterogeneity

What if it’s not actually possible to make all people to look more similar to the typical high earning people? Maybe some things in our data are impossible to change. In this case, it’s

---

18. If $P(X = x|I = 0) = 0$ (which means that $x$ is not observed in the data), then $P(X = x|I = 1) = 0$, otherwise $D_{KL} = \infty$.

19. We also found that higher values of $\lambda$ yield results more similar to ones we get under linearity assumptions.
more sensible to try to make people to be more similar to people that have higher income but are more similar to them on other dimensions. In the language of our example, rather than making $\theta_0$ more like $\theta_1$, we can make the $\theta_0$ who don’t graduate from college more like the $\theta_0$ who do graduate.

To put this in the context of our framework, we assume that there is heterogeneity in the distribution of $X$. Hence, it’s possible that high earning people are drawing $X$ from a different distribution. This would mean that drawing such $X$s could be much more costly for some people. Formally, we will assume that $P_i(X|I=0)$ is different for every observation $i$. Moreover we will assume that this probability is higher for neighboring values: other values of $x$ we observe in the data for observations that are close. As a result, the cost $C_i(I)$ is lower when we try to change $x_i$ to its neighbors value.

**Assumption 2.3.** Every individual $i$ draws $X$ from the following distribution:

$$P_i(X = x|I = 0) = \begin{cases} 
\exp\left(-\frac{\text{dist}(x,x_{i'})^2}{2\sigma^2}\right) & \exists x_{i'} \in \text{data s.t. } x_{i'} = x \\
0 & \exists \hat{x}_{i'} \in \text{data s.t. } x_{i'} = x
\end{cases}$$

and $\text{dist}(x,x_{i'})$ is Mahalanobis.

The parameter $\sigma$ will set how much we penalize for distance. When $\sigma \to \infty$ we can turn child $i$ to any other child in the data with equal costs, and so there’s no heterogeneity. As $\sigma \to 0$, we can only change to the closest neighbor. There is a tradeoff between bias and variance in the choice of $\sigma$. High $\sigma$ will use all data, and would therefore be more biased but with less variance. Low $\sigma$ will use fewer and closer data, and will therefore be less biased but noisier. We use $\sigma = 1$, but different values yield similar results. For the choice of $f$ we will use the same non parametric method we used before and set $\hat{f}(x_i) = \log w_i$.

As we’ve seen before, an intervention $I$ that sets $P_i(X_i = x_0|I = 1) > 0$ when $x_0$ is not
in data will make the KL divergence infinite. Therefore we can limit ourselves to inter-
vention that sets some positive probability $P_{ij} = P_i (X = x_j | I = 1)$ where $x_j$ is the $j$th
observation in the data, and $P_i$ is the specific probability distribution of the $X$s for child
$i$. Our goal is then to choose values for $P_{ij}$ for every $j$ s.t. $\sum_j P_{ij} = 1$.

**Proposition 2.4.** Assume $P, f$ are distributed as above and Assumptions 1 & 2 hold. Then for
the optimal choice of $I$

$$P_{ij} \propto w_j^{\lambda - 1} \exp \frac{\text{dist} (x_i, x_j)^2}{2\sigma^2}$$

Intuitively, this exercise is similar to the reweighting in the non-parametric section. The
key difference is that now we put more weight on closer neighbors. Therefore, high-
income people who look very different from the rest of our data would get a lower weight,
compared to the non-parametric case. This captures the intuition that people that look
very different, might have their $X$s produced in a different way, and therefore an inter-
vention that would try to make all people more similar to them would be less likely to be
successful. Overall, under $I = 1$ we would have a distribution of $X$ with the same sup-
port of our data, that has a higher probability to draw $X$ values of high income people,
who are similar to the $X$ distribution in the data.

We choose $\lambda = 100$ as we chose in previous sections. We calculate $P_{ij}$ up to a constant,
using the above equation, and normalize to get the probabilities sum to one. This gives
us a distribution to draw each value of $x_j$ which is $P (X = x_j | I = 1) = \frac{1}{N} \sum P_{ij}$. Using
this distribution, we calculate the expectation of each variable, and compare it to the
expectation in the data ($I = 0$), as we did in previous methods.
2.4. Results

In this section, we present results gleaned from implementing the methods described above on our new set of data. Each one of the methods yields a new distribution of $P(X|I = 1)$, which we compare to the original distribution of the data $P(X|I = 0)$. Our primary goal is to detect which variables undergo the biggest changes. We do this by plotting a series of figures with the change for each variable in rank order, for all variables in our dataset that are statistically different from zero.

Formally, for every variable $X_j$ we first calculate

$$\tau(X_j) = \frac{|E[X_j|I = 1] - E[X_j|I = 0]|}{\sqrt{V(X_j|I = 0)}}$$

where $\tau(X_j)$ is the absolute difference in expectations of a variable under $I = 0$ and $I = 1$. We divide by $\sqrt{V(X_j|I = 0)}$ to make this unit-free.

Second, to make the difference in expectations comparable across methods, we standardize the distribution of $\tau(X_j)$ i.e. we divide $\tau(X_j)$ by its standard deviation across all variables in a given method:

$$\frac{\tau(X_j)}{\sqrt{\sum_{j'} \left( \tau(X_{j'}) - \overline{\tau(X_{j'})} \right)^2}}$$

One can interpret the x-axis in each figure as the units of standard deviation for $\tau(X_j)$. The bars surrounding each coefficient estimate is the 90% confidence interval.
Figure 2.2.: Non-Parametric Method
2.4.1. Income

We begin with log household income as our outcome variable and our preferred specification – non-parametric estimation of $P$ and $f$ (Figure 2.2a). The most important correlate of income mobility is education. This is consistent with a large literature on the importance of the quantity of education on income (Card, 1999; Garces et al., 2002; Belfield, 2006; Barnett and Masse, 2007; Heckman et al., 2010, 2013; Elango et al., 2015; Heckman et al., 2015). A close second – and statistically indistinguishable – is resilience. Recall, resilience is the ability to bounce back from stressful situations and is measured by responses to questions such as “It does not take me long to recover from a stressful event”.

Surprisingly, half of the significant correlates of intergenerational income mobility are psychological skills: resilience, Big 5, self-esteem, self control and grit. Other important variables are whether the respondent was ever in trouble with the police in their youth, had adverse childhood experiences, and the number of adult relationships they trusted.

If we adjust household income by household size, we get similar results though estimates are noisier (Figure 2.2b).\(^20\) If we use individual income as our measure, rather than household income, we obtain a larger set of significant variables (Figure 2.2c). All the variables above are significant along with risky attitudes as a teenager, whether the respondent lived with a mother when they were young, and growth mindset – another psychological skill.

These results are in contrast to much of the literature on the correlates of income mobility, though consistent with Nyhus and Pons (2005); Heckman et al. (2006); Currie and Spatz Widom (2010); Moffitt et al. (2011); Heckman et al. (2013). For instance, we do not find that church going, fraction fathers present in a zipcode, or mobility indices more generally are significant correlates. Generally, there is a larger focus in our results on psy-

\(^{20}\) We adjust income for household size using the same method employed to calculate the Census Supplement of Poverty Measure (Fox, 2018). A detailed description of the adjustment is provided in the Appendix B.3.
Chological skills and the ecosystem embodied by children when they are young, which includes interactions with police and other adverse childhood experiences, risky behaviors, and the adults in a child’s life.

Figure 2.3 shows the same results assuming that \( X \) is multivariate normal and \( f \) is linear. This is equivalent to partial correlations, controlling for parents income. The Spearman (rank) correlation between the non-parametric method of estimating \( P \) and \( f \) and assuming that \( X \) is multivariate normal and \( f \) is linear is .90. But because of the parametric assumptions we have more statistical power which leads to more significant variables. Again, under these assumptions, we get all the variables that we did above plus mental health before the age of 18, locus of control, family environment, and fraction of fathers present in a zipcode. Similar to above, of the fourteen variables that are significant, seven of them are specific psychological skills.

Figure 2.4 shows the results when we allow for heterogeneity. The correlation between the non-parametric approach and when we account for heterogeneity is .91. All but one variable from the non-parametric approach continues to be significant after we allow for heterogeneity and they are the highest ranking correlates (the exception is number of adults trusted). Because of smaller standard errors, when we allow for heterogeneity we have the following additional correlates: family environment, parenting, family network, risky attitudes as a teenager and three additional psychological skills (locus of control, growth mindset, and self control).

Different methods offer modest differences in the specific variables that are gleaned to be significant. But the general pattern is robust. Education, psychological skills, trouble with the police, and adverse childhood experiences are always significant, independent of method. This is consistent with recent evidence that education centric interventions designed to increase income among the poor – so-called No Excuses charter schools – may have little impact on mobility (Dobbie and Fryer, 2016). And, suggests that these
Figure 2.3: Partial Correlations

(a) Household Income
(b) Adjusted Income
(c) Individual Income
Figure 2.4: Nearest Neighbor
interventions would be more successful if one simultaneously worked to increase psychological skills and the number of adults trusted, and reduce trouble with the police and other adverse childhood experiences.

2.4.2. Analysis of Subsamples

Appendix Figures B.6-B.11 explore the sensitivity of our income correlates across a variety of subsamples of the data. We report our estimates that assume $X$ is normal and $F$ is linear separately by race, gender and parental income. We opt for this specification given its high correlation with the non-parametric approach and smaller standard errors.

For each division we plot the variables that are significant for both groups, and the variables that are significantly different between groups (both at the 90% level). For race, we find that two variables are significantly different between blacks and whites. Consistent with Chetty et al. (2018) we find that high mobility zipcodes are significantly more helpful to whites. We also find that number of adults one could trust and whether one could trust any adults in their childhood is operating in opposite directions for blacks and whites; while it is increasing black’s income, it is decreasing in white’s income. Although we are not certain about the mechanism that drives this difference, we see that for all respondents that said “yes” to trusting adults in their childhood, whites had a significantly higher fraction of unemployed fathers than blacks. So white respondents may have depended on fathers that were untrustworthy which could have resulted in lower adult incomes.

With subsamples based on gender, we find that family network seems to be much more important for girls. Other variables have statistically similar impacts on boys and girls.

Finally, we split our sample by parents income. We find that two variables related to parents are going in opposite directions. For people who grew up in deeper levels of
poverty, income is higher for worse parents behavior.

2.4.3. Adult Well-Being

Thus far, we have concentrated on variants of household and individual income as outcomes. Even for economists, however, there is more to life than income. In this section, we explore a wider definition of adult well-being by including adult physical health, mental illness, and alcohol and drug abuse. Figure 2.5 shows our results for three additional outcomes that we will now discuss.

*Mental Illness*

The correlates of mental illness are quite different from those correlated with income. Unsurprisingly, six of the top eight correlates of adult mental health are psychological skills measured in adulthood. It is unclear what this means. Ideally, the psychological skills would be similar in childhood and thus interventions on those variables may prevent adult mental illness. It is also plausible however, that whatever is associated with adult mental illness is also associated with lower values of general psychological skills.

More interesting, risky attitudes as a teenager and mental health as a teenager are also associated with adult mental health. Education, physical illness in childhood and adverse childhood experience are also associated with adult mental health in the expected directions. Fathers present in zipcode, whether the respondent lived with their father as a child, trouble with police, family environment, and neighborhood safety are also all significant correlates of adult mental health. The similarity between these results and those from the Moving to Opportunity experiment are striking. In that experiment, moving poor individuals to less poor neighborhoods was associated with significant improvements in mental health.

*Alcohol and Drug Abuse*
(a) Adult Mental Illness

(b) Drug and Alcohol Use

(c) Adult Physical Health

Figure 2.5.: Non-Parametric Method, Alternative Outcomes
The correlates of alcohol and drug abuse are interesting and intuitive. The most important correlates are risky attitudes as a teenager and self control, followed by trouble with police, mental health before sixteen years of age, and grit. Other variables include various parenting variables and family environment and neighborhood safety. It is important to note, education is not in the top 10 of most important variables. The variables above are more important if one wants to reduce drug and alcohol use.

Physical Health

The top correlates of adult physical health are surprisingly intuitive. The variable physical illness before 16 is now among the top variables, as there is some correlation between physical illness in childhood and adulthood. Good diet is also associated with better adult health. Just as with adult mental illness, adult physical health are also highly correlated with psychological skills, and childhood mental illnesses. Other top correlates are education and risky attitudes as a teenager.

2.4.4. Analysis of Subsamples

Observing the differences of correlates by race we see that, similar to household income, trusting adults during childhood leads to better mental health for black kids but worse mental health for white kids. Living with mother or stepmother during childhood leads to worse mental health outcomes for black kids. This may be an indicator of worse outcomes for black kids who have grown up with absent fathers, a result consistent with Chetty et al. (2018). The only other variable that is statistically different for blacks and whites and which affects an outcome in opposite directions is family network. For black children, strong family networks lead to worse adult physical health while for white kids, a strong family network leads to better adult physical health. Without more knowledge of how one’s network looks like, it is difficult to surmise the reason behind this result.
Splitting subsamples by gender, we observe that all correlates impact adult mental health in the same direction for both boys and girls. If anything, girls’ mental health seems to be impacted to a higher extent than boys’ mental health for any given correlate. With regard to adult drug and alcohol use, the only surprising result that we have which is difficult to explain is that girls with higher beliefs about future success seem to have higher drug and alcohol use. The correlate is negative and insignificant for boys. For adult physical health, we see that living with a grandparent might have a negative impact on future physical health for boys. The same result does not apply to girls.

Our final subsamples are obtained by splitting across parental income. In general, just as with household income, we see that children with lower income parents require stronger intervention for adult mental health and adult drug and alcohol use.

### 2.5. Discussion

Our analysis of intergenerational mobility has developed a new set of facts. Using newly collected data and a new method, we argue that to increase income among the poor we need a multi-pronged strategy that focuses heavily on the ecosystem children inhabit, their psychological skills to navigate the situations they endure, and, importantly, education. These results are suggestive. We caution against a rush to policy, but rather a rush to experimentation with the goal of boosting income among those who are born poor.

Our analysis has several important caveats.

First, our locations are not representative. Yet, as we argued in section 2.2, our sample of respondents who grew up poor in Memphis, Tulsa and New Orleans look similar to a nationally representative sample gathered from the 2014 wave of NLSY who were between the ages of 14 and 22 and below the poverty line in 1979.
Second, our data points are retroactive. We ask respondents to recall their home environment, parenting details, and experiences from childhood, many years removed. Not only is memory unreliable, but there may be systematic biases in the types of narratives individuals believe about their life trajectory. That is, individuals who are more upwardly mobile may emphasize the highlights of their childhood experience whereas individuals who are less successful may emphasize the lowlights. Assumption 1 dealt with this issue theoretically. But, the empirical importance of these potential biases is untestable with current data.

In addition, our variables designed to assess psychological capital – grit, growth mindset, resilience, locus of control, Big 5 personality traits, self esteem and self control – are contemporaneous. It is possible that respondents who have escaped poverty feel more resilient or have higher self esteem. This is partially testable.

To understand how this might affect our results, we analyze a longitudinal dataset where respondents are queried about their psychological traits in multiple waves. The NLSY assesses respondents’ locus of control in 1979 and then again in 2014. We estimate the correlation between income in 2014 and locus of control in 2014 – which is similar to what we do in our analysis – and then separately, estimate the correlation between income in 2014 and locus of control in 1979. If the correlation coefficient from both specifications are statistically similar, it provides some evidence that an internal locus of control has a positive impact on adult income regardless of when this locus of control was measured.

Appendix Figure B.12 presents the results. Similar to our data, locus of control is positively correlated with income. The slope using our dataset is 0.07 (0.03) and 0.12 (0.02) in the NLSY data when regressing current income on current measures of locus of control, both have p-values below conventional levels. Importantly, when we plot current income on locus of control in youth using the NLSY data, the slope is 0.05 (0.02) and is also significant. The p-value of the difference in coefficients on locus of control from the 3 plots
is 0.158. This shows that the correlation between income and locus of control from our dataset is statistically similar to the correlation between income and locus of control from the NLSY sample regardless of when locus of control is measured in the NLSY sample. Third, our results are best seen as a list of “sufficient” – not necessary – variables to increase a particular outcome. They are correlates, not causal estimates, though some of the variables have been shown in other work to have a causal relationship. Appendix B.8 summarizes the literatures that estimate the causal relationship between variables we believe our significantly correlated with income, separately. We do not have estimates of the total derivative, but there are several important partial derivatives. It is a reasonable exercise to believe that as long as $X_1$ is positively causally related to $Y$ and $X_2$ is positively causally correlated with $Y$, then as long as $X_1$ and $X_2$ are positively correlated with each other, we would expect an intervention in both $X_1$ and $X_2$ to increase $Y$. Appendix Table B.7 already shows that the most important correlates with household income are positively correlated with each other, so positive partial derivatives from literature would be a good guess about positive total derivatives from potential interventions.

Years of education has a causal impact on income. Card (1999) gives an incredible survey of literature on causal relationships between education and earnings. He estimates that an additional year of schooling can increase wages by 2-11%. Heckman et al. (2015) study the subsample of males extracted from NLSY 1979 to estimate the positive increase in earnings caused by an additional year in high school or college. Besides education in higher grades, there is also a vast amount of literature on the efficacy of early childhood education on earnings. Belfield (2006); Heckman et al. (2010, 2013) measure the positive impact of the HighScope Perry Preschool Program on participants’ earnings. In the same vein, Barnett and Masse (2007); Elango et al. (2015) calculate the net effect of the Abecedarian program on adult earnings and Garces et al. (2002) and Elango et al. (2015) calculate the same estimate for Head Start participants. Appendix B.8 gives details about
these papers and the corresponding causal estimate calculated by each.

There also exists a diverse set of papers studying the impact of higher scores on the Big 5 Personality Traits on earnings. Nyhus and Pons (2005) study 888 workers aged 16-65 who were part of the CentER Saving Survey in Netherlands and conclude that higher scores on different personality traits like agreeableness and emotional stability leads to higher wages for men and women. Heckman et al. (2013) study 123 HighScope Perry Preschool Program participants and infer that their increase in adult incomes is causally due to improvements in Big 5 personality traits.

Heckman et al. (2006) also estimate the effect of higher self esteem and more internal locus of control through their study of NLSY 1979 participants and conclude that a 1 standard deviation increase in non-cognitive ability increases hourly wages by 11.2%. Moffitt et al. (2011) study sibling pairs from the Dunedin Multidisciplinary Health and Development Study in New Zealand and state that higher self control leads to higher adult incomes. Currie and Spatz Widom (2010) track 807 individuals from a Midwestern metropolitan county which include children who have court substantiated cases of childhood physical and sexual abuse and neglect and study their labor market performance compared to a matched sample of non-abused and non-neglected children. They estimate that child maltreatment reduces adult earnings by approximately $5000.

Finally, there is also literature on the causal impact of youth criminal behavior on adult income. Allgood et al. (1999) analyze 439 sibling pairs from NLSY 1979 and estimate that a criminal charge when young leads to a reduction of adult income by 22% while a criminal conviction reduces adult income by 36%.

This type of analysis is, at most, speculative. Whether the bundle of variables we view as significant correlates can increase income is an experimental question. Will it work? We can’t know in the abstract. But, with the benefit of hindsight, we can use our method on
the detailed within-school data collected in Dobbie and Fryer (2013) and then compare the suggested optimal intervention from our to the intervention implemented in Fryer (2014).

Dobbie and Fryer (2013) study data collected from 39 charter schools and correlate it with estimates of school effectiveness. They find that traditionally collected input measures – class size, per pupil expenditure, fraction of teachers with no certification, and the fraction of teachers with advanced degree – are not correlated with school effectiveness. In contrast, policies suggested by qualitative research – frequent teacher feedback, the use of data to guide instruction, high-dosage tutoring, increased instructional time, and high expectations – explains 45% of the variation in school effectiveness.

Using the same data, we implement our method. The results are interesting and are shown in Appendix Figure B.13. The significant correlates of high quality schools are teacher feedback, instructional time, high expectations and high quality tutoring. Interestingly, non-certified teachers are strongly negatively correlated with school effectiveness while data driven instruction is not significantly correlated with school effectiveness.

Fryer (2014) implemented 4 of these and demonstrated large impacts in math and less in reading, a pattern closely resembled in the achievement-increasing charter schools they were gleaned from. Whether Fryer (2014) would have had significantly different results if they had removed non-certified teachers as the suggested correlates from our method shows, is unknown. However, it is important to note that given that literature suggests that certified teachers, better teacher feedback, higher instructional time, high expectations and high quality tutoring have positive partial derivatives with respect to achievement and the data suggests that they are positively correlated with each other, we may reasonably expect an intervention using the suggested correlates from our method to have a positive effect on math and reading scores.
Overall, as a discipline, we have strong evidence that intergenerational mobility fairly low for some population. More importantly, we also know now that it is malleable and that differences in childhood environments could lead to substantial differences in adulthood. Therefore, there is large potential for improvement. However, we know very little about which aspects of childhood environment are more important than others. The next step in this literature is to run more randomized experiments to pin down a cost effective way to increase mobility. Given the difficulty and costs in running such experiments, it is highly important to prioritize experiments with higher likelihood of success.

This paper uses new data and new methods to argue that our focus should be mostly on education and psychological characteristics. We find that it is much more common to escape poverty with better education and better psychological characteristics than it is with better neighborhoods, or better physical health for example. Therefore we see greater potential in a policy that would intervene in such characteristics, and hope to see experiments that are testing this.
3. Decomposing Wage Polarization in the U.S.

3.1. Introduction

During the decade of the 1990s, the U.S. labor market has experienced wage polarization - a substantial relative decline in middle wages. Wages around the median declined compared to both wages at the top and at the bottom. Figure 3.1 show this trend was different from the broad increase in inequality at all parts of the distribution that occurred both before, in the early 1980s, and after, starting from the early 2000s. The leading explanation for this trend is a Routine-Biased-Technological-Change (RBTC): a decline in demand for routine tasks that used to require middle-skill workers, but now can be automated (Autor et al., 2006). This hypothesis is supported by the extensive decline in employment at routine-heavy occupations in most developed countries (Goos et al., 2014).

But while there is a variety of strong evidence that support the mere existence of RBTC, very little is known about whether the magnitude and the shape of its effect on wages can explain large portions of the decline in middle-wages. It is also harder to explain why RBTC is mostly affecting middle-wage workers, where many routine workers are actually concentrated at the bottom of the distribution. Another puzzle is why the decline in
middle wages stopped in mid-2000s when evidence suggest RBTC continues long after that. These unresolved questions left room to consider other explanations to wage polarization such as an increase in minimum wage (Piketty, 2014), decline in unions (Firpo et al., 2013), business cycles (Foote and Ryan, 2015), demand growth in the service sector (Autor and Dorn, 2013) or the low unemployment rate during the 1990s.

One reason why we know so little about the relative contribution of each explanation to the general trend is that we do not have the right tool to decompose wage polarization. While there is quite a diverse toolkit to decompose wage inequality, wage polarization possess new challenges that need to be addressed. The key challenge is that, by definition, wage polarization has asymmetric trends as the wage gap increases at the upper tail of the wage distribution, but decreases at the lower tail. Most existing decomposition methods can isolate the effects of prices and compositional changes in the overall trends in wages. But if prices changed in opposite directions at different parts of the distribution, this will

**Figure 3.1:** All Population 90/50 and 50/10 Log Hourly Wage Ratio
Quantiles are calculated at the hours level, using sample weights multiplied by hours worked.
Source: CPS Outgoing Rotation Groups
not be captured.

This paper will show that RBTC does in fact account for almost the entire trend of wage polarization. I will do so using a decomposition method which I call “Skewness Decomposition”. Using this method I estimate that 79% of wage polarization are due to asymmetric trends in occupations: a rise in inequality at high paying occupations and a decrease at low paying occupations. Both of these trends are generating a relative drop in middle wages. But while rising inequality at high paying occupations has been steady for several decades, the period of the 1990s is unique for its decline in inequality at low-paying occupation. Using panel-data I show that this is driven by a wage compression at routine occupations.

Using a simple model, I show how RBTC can explain these findings, and address several puzzles about the effect of RBTC on wages. The key distinction from previous models is that the new technology is not equally substitutional to all workers in routine tasks. Instead, new technology is substitutional to the usage of skills in those tasks. This generates a decrease in return to skills, only in those heavily-routine occupations. Even though most routine workers are earning below median wages (Autor and Dorn, 2013), such a decrease in return to skill is affecting mostly middle-wage workers, because that’s the typical wage level for the higher-skilled workers in routine occupations. As a result, higher-skilled workers leave these occupations, gradually making routine occupations composed of more low-skilled workers, without many middle-wage workers. From that point, any further RBTC is generating a decrease in demand for low-skill workers, and inequality starts increasing again at the lower tail of the distribution.

I start by outlining the theoretical framework of the paper. The key assumptions is that workers are characterized by a one dimensional skill, and occupations vary in their returns to that skill. This means that workers with a higher skill level will have a comparative advantage in occupations with a higher return to skill. Occupations with mainly
manual tasks have the lowest return to skill, and occupations with mainly abstract tasks have the highest return. Occupations with mostly routine tasks provide comparative advantage to workers from the middle of the skill distribution. In the empirical section, I test these assumptions and show that they work reasonably well in the data.

The key parameter in the model is the elasticity of substitution between the skill and the new technology. The model is general enough to allow for a technological change that is skill neutral, as in many previous models (i.e., Acemoglu and Autor, 2011; Cortes, 2016). It also allows for a technological change that is increasing returns to skill as in Jung and Mercenier (2014), or decreasing as causal evidence suggests (Gaggl and Wright, 2017). The model will allow us to derive different predictions for each one of these cases, which we will then be able to test in the data.

In order to quantify the overall effect of RBTC on the wage distribution, we need to use some decomposition method. Decomposition methods are the standard way to analyze changes in distributions across time, or between groups (Fortin et al., 2011). For instance, using variance decomposition, it is very straightforward to see that most of the increase in income inequality during the 80s was through increasing gaps between education groups (Katz and Autor, 1999). To study the effect of RBTC, we would like to decompose by occupations, as it is the key component of the underling theory behind it.

However, there are various reasons why existing decomposition methods would not work. Since wage polarization could both increase or decrease inequality, methods that are designed for inequality analysis would be irrelevant. Instead we need a new statistic that measures polarization, that could be decomposed. More general decomposition methods, that simulate full counterfactual distributions (Juhn et al., 1993; DiNardo et al., 1996) or that could decompose any statistic (Firpo et al., 2009) also pose a different set of problems. Generally, they are harder to interpret due to reasons such as path-dependence or an arbitrary choice of baseline year (Fortin et al., 2011).
But more specifically, these methods are not suitable to capture asymmetric trends. Broadly speaking, most decomposition methods are decomposing any change in wages to change in composition, prices and a residual. In the context of occupations, they can capture any wage trend that is driven by transitions between occupations, or change in mean wages at each occupation. But as the model predicts, and as the results suggest, a large share of the trend in wages is through asymmetric effect on inequality within occupation. This will be missed by those methods that will classify it in the residual component.

Skewness decomposition can solve all of those problems. In analogy to inequality that can be measured with the second moment of the distribution of log wages, polarization can be measured with the third moment of that distribution - the skewness. As expected, the skewness is rising exactly when wage polarization is increasing. The advantage of using skewness is that, similar to variance, it can be decomposed into independent components, that do not depend on any arbitrary choice of year or order of components.

Skewness decomposition breaks the trend in wage polarization into three components, for each choice of groups. The researcher can choose a group such as education level, occupations, industry etc. and decompose the increase in polarization by that group. Similar to variance decomposition, skewness decomposition has a between-group and a within-group components. The between-group captures any trend that is driven by changes in prices between the groups. The within component captures any unexplained trend that is orthogonal to that grouping. But there’s an important third component that captures the correlation between the mean and variance at each group. This component will be higher when higher paying groups have larger inequality. This allows skewness decomposition to capture some types of asymmetric wage changes, that are missed in other methods.

Applying skewness decomposition to data on the wage distribution in the U.S. gives a direct evidence that wage polarization is driven by occupational trends. So far, the main
suggestive evidence that linked wage polarization to occupations was employment and wage decline in routine-heavy occupations during the same period of time (Acemoglu and Autor, 2011; Cortes, 2016). Those routine occupations tend to pay low to middle wages, suggesting that this drop in demand could be linked to wage polarization. My results quantify the share of the rise in skewness that can be attributed to occupations and show it can explain 93% of the overall increase. Comparing these results to the result when decomposing by other categories, such as industries or education, shows clearly that the trend is driven by occupations. I use data from the CPS outgoing rotation group since it measures the price of labor most precisely (Lemieux, 2006).

But in contrast to the prediction of earlier models for RBTC, the effect is driven by asymmetric inequality trends within occupations. Earlier models (Acemoglu and Autor, 2011) assume that all routine workers have the same skill level, and that RBTC simply changes the price of routine tasks. This setting predicts that wage changes would be made through the decline in premium for routine occupations, and doesn’t allow for any distributional changes within routine occupations. If that was the case, skewness decomposition should have shown that the increase in skewness is in the “between” component. However, I find that almost the entire increase is in the correlation component. Wage polarization happened because inequality increased at high paying occupations, but decreased at low-paying occupations. This trend was documented by Lemieux (2007), and using skewness decomposition I find it is actually the main driver of wage polarization, far beyond the changes in occupation premiums.

Wage polarization is driven by the drop in inequality in heavily routine occupations. While inequality at high-paying occupations is steadily increasing for decades, the drop in inequality at low paying occupations is unique to the 1990s. This is the reason why wage gaps are falling at the bottom of the distribution, generating wage polarization during that period. Most of this drop is in low-paying routine occupations, while other low
paying occupations like services don’t experience any such trend.

There are two distinctive explanations for this trend, and I use panel data to decide between them. One reasonable explanation is that the drop in demand for workers in routine tasks made the highest and lowest paid workers to leave, thus making the distribution more equal. I show the model generates this pattern when RBTC is skill neutral. Cortes (2016) shows empirical evidence that generally workers from the edges of the wage distribution are more likely to leave, so potentially this trend exacerbated during the period of wage polarization. An alternative explanation is that returns to skills have declined, generating a wage compression even without any transitions. To distinguish the two explanations I use panel data from the PSID.

I find that a decrease in returns to skill at routine occupations in the key driver of wage compression in routine occupations, and hence of wage polarization. I estimate an interactive fixed effect model, that allows the return to skill to vary across time between occupations. I find evidence for a decrease in return to skill in routine occupations. This decrease in routine occupations makes wages to drop mostly for the highest earning routine workers, who are concentrated around the median of the overall wage distribution. As a result, higher-skilled workers have the highest incentive to leave routine occupations. Indeed, I find that almost all of the employment decline in routine occupations (“job polarization”) is driven by workers with above average skills. The share of below average workers in routine heavy occupations remains steady.

These results explain how RBTC can first generate a decline in middle wages, which then stops and turns into a decline in lower wages. At first, there are several high skill workers in routine occupations. Because wages in routine occupations are relatively lower, those workers’ wages are close to the median wage of the entire labor market. The drop in demand is strongest for those workers, who experience the largest drop in their wages and eventually leave. This generates both wage and job polarization. Gradually, the
composition of routine occupations becomes more affluent with low skilled workers. Any further decline in demand for routine tasks would then be a drop in demand for lower-wage workers and generate a decline in lower wages.

Other explanations cannot explain these findings. Institutional explanations like an increase in minimum wage or decline in unionization, as well as macroeconomic explanation like low unemployment do not fit these empirical findings. In general, these explanations are not particularly related to occupations, more than they are to industries, education levels or other partitions of the data. Moreover, they shouldn’t affect workers in routine occupations any differently from workers in services or in more abstract-heavy occupations. Increase in demand for service occupations seems to be more of a result of polarization and not the main driver of it. If it were, we would expect most of employment decline to be driven by lower-skilled workers, which is the opposite of what I find.

The rest of the paper is organized as follows: Section 3.2 outlines a theoretical framework for RBTC. In Section 3.3 I will discuss skewness decomposition in detail and its advantages over existing methods. Section 3.4 describes the data sets used throughout the paper. Section 3.5 will present the results of skewness decomposition and show wage polarization is driven by occupational trends, supporting the RBTC hypothesis. Section 3.6 will present the evidence for a decrease in returns to skill at routine occupations, using panel data. Section 3.7 will conclude by reexamining all the evidence and show how they fit the model predictions.

3.2. Model

In this section I’ll present the theoretical framework for RBTC that will be used thought the paper. I use a model that highlights the different return to skill in each occupation, in
the spirit of Jung and Mercenier (2014) and Cortes (2016). The model allows to introduce new technology that could be substitutional, complimentary or neutral to the workers’ skill. In each of those cases technology will affect the distribution of wages within occupation, between occupations and the composition of workers in each occupation, but in different ways. I’ll discuss the differences in these predictions and how they can be observed in the data.

3.2.1. Simple Model of Occupational Sorting

Assume workers have a one dimensional skill. We will mark this skill level by $\theta_i$. This assumption is more general than models that assume only a discrete number of skills (Katz and Murphy, 1992; Autor et al., 2006; Acemoglu and Autor, 2011), but less general than models that allow for multi-dimensional skills (Roy, 1951). I will test this assumption empirically in Section 3.6, and show that adding more dimensions of skills doesn’t improve precision substantially.

Occupations will be characterized by their return to skill. To simplify I will assume three occupations: manual, routine and abstract. In each occupation $j \in \{M, R, A\}$ workers produce an intermediate good with a production function $\varphi_j(\theta_i)$. Assume that

$$\forall \theta : \frac{\partial \log \varphi_M (\theta)}{\partial \theta} < \frac{\partial \log \varphi_R (\theta)}{\partial \theta} < \frac{\partial \log \varphi_A (\theta)}{\partial \theta}$$

(3.1)

so the manual occupation has the lowest usage of skill, and abstract has the highest.

Perfect competition sets wages at their marginal productivity. I assume that identical firms are competing for the same workers. Let $p_j$ be the price of the intermediate good in occupation $j$. Therefore, if worker $i$ is working in occupation $j$, she will earn

$$w_j(\theta_i) = p_j \varphi_j(\theta_i)$$
Workers will sort into occupations based on comparative advantage. Condition 3.1 guarantees that there would be two thresholds $\theta_0, \theta_1$ such that any worker with $\theta_i < \theta_0$ will choose to work in manual, any worker with $\theta_0 < \theta_i < \theta_1$ will choose routine, and $\theta_i > \theta_1$ will choose abstract (Jung and Mercenier, 2014). Workers with skill level that exactly equals the threshold will be indifferent, hence the following two equations will hold in equilibrium

$$p_M \varphi_M (\theta_0) = p_R \varphi_R (\theta_0)$$

$$p_R \varphi_R (\theta_1) = p_A \varphi_A (\theta_1)$$

Figure 3.2 presents this graphically, by plotting the equilibrium log-wages by skill level $\theta_i$. 

Figure 3.2: Equilibrium log Wage by Skill
3.2.2. Technological Change

I assume that the three intermediate goods, together with capital are producing a final good. Mark the total amount produced from each intermediate good by $M, R, A$ with

$$
M = \int_{\theta_{\min}}^{\theta_{1}} \varphi_{M}(\theta) \, d\theta \\
R = \int_{\theta_{0}}^{\theta_{1}} \varphi_{R}(\theta) \, d\theta \\
A = \int_{\theta_{1}}^{\theta_{\max}} \varphi_{A}(\theta) \, d\theta
$$

and assume total capital $K$ is fixed.\(^1\) The final good is the output of a CRS function $Y = F(M, R, A, K)$.

I focus on a technological change that is shifting the production of routine goods from labor to capital. I assume that technological change is affecting only $\varphi_R$ directly, and $\varphi_M, \varphi_A$ are left unchanged for simplicity. However, wages in the manual and abstract occupations will be affected as well in general equilibrium. Specifically I assume the following functional form

$$
\varphi_R(\theta_i) = \left(\theta_i^{\frac{\sigma-1}{\sigma}} + \tau\theta_i^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}
$$

where $\tau$ is the level of technology. $\sigma$ is the elasticity of substitution between technology and skills.

Capital is complimentary to routine work, so if routine workers produce more, a larger share goes to capital. To simplify, I assume a specific functional form to $F$ where

$$
Y = M^{\alpha_1} (R^{\rho} + K^{\rho})^{\frac{\alpha_2}{\rho}} A^{1-\alpha_1-\alpha_2}
$$

and $\rho < 0$. The Cobb-Douglas structure implies that a constant share of $Y$ goes to manual

\(^1\)The results could be generalized for a case where capital has increasing marginal costs.
workers, and abstract workers. A constant share of $\alpha_2 Y$ goes jointly to routine workers and capital owners.

RBTC would then be modeled as an increase in technology level $\tau$. This allows for every worker to produce more, since $\frac{\partial \varphi_R(\theta, \tau)}{\partial \tau} > 0$. Therefore, RBTC will generate an increase in $R$, which would increase the share going to capital from total output on the expense of total share to routine workers, since they are complements. An increase in $\tau$ can be thought of as an improvement in quality of computers or robots. While this makes each worker more productive, it also allows to produce the same quantity with fewer workers, making labor prices drop.

To understand the effect of RBTC on the labor market, it’s key to know whether the elasticity of substitution between technology and skills ($\sigma$) is greater, equal, or less than 1. The effect of an increase in $\tau$ on different levels of $\theta$ depends on

$$\frac{\partial^2 \log \varphi_R(\theta, \tau)}{\partial \theta \partial \tau} = \frac{1 - \sigma}{\sigma} \frac{1}{\left(\frac{\theta^{\frac{1}{\sigma}}}{\theta^{\frac{1}{\sigma}}} + \frac{\tau^{\frac{1}{\sigma}}}{\frac{1}{\tau^{\frac{1}{\sigma}}}}\right)^2 (\tau \theta)^{\frac{1}{\sigma}}}$$

which has the same sign as $1 - \sigma$. If $\sigma$ is 1, RBTC is skill neutral as in Cortes (2016). The effect on log wages will be the same for all workers in routine occupations. If $\sigma < 1$, as hypothesized by Jung and Mercenier (2014), the new technology is increasing gaps between skill levels. At some point, routine occupations could actually have a comparative advantage for the highest skill workers. If $\sigma > 1$, technology is substitutional to skills, and returns to skill will decline.

In any case, simple decomposition methods will not capture the entire effect of RBTC. If $\sigma \neq 1$, RBTC is not only changing the premium for routine occupations, it also changes

---

2A more general CES function with complementarities between occupations (as empirical evidence suggest) could be used instead, and the results will only be amplified. In case of complementarities, RBTC that allows for higher $R$ will cause larger share of total income to go to manual and abstract occupations. This will increase the decline in routine wages. For simplification we’re using CD.
wage gaps within routine occupations. Even if $\sigma = 1$, the distribution of wages within occupations will change. The lowest (highest) skill workers will leave to manual (abstract) occupation. This will decrease inequality at routine occupations, and increase it in manual and abstract occupations. Any such changes are not captured by a decomposition method that is focusing on prices and compositions.

3.2.3. Decrease in Return to Skill

I will focus in more details on the case where technology and skills are substitutional, as it will have the highest fit to my empirical results. I will describe the different stages of a gradual increase in $\tau$, and how they can be measured empirically.

At first stage, we would see wage polarization. Wage polarization will be generated by a decrease in inequality in routine occupations. Figure 3.3 illustrates this case where the return to skill in routine occupations becomes flatter. This generates lower gaps between workers who stay at routine occupations. The most significant wage drop is for workers at the top wage levels of the routine occupation, which are approximately in the middle of the overall distribution of skills. As a result, higher skill routine workers will now have their comparative advantage in the abstract occupations, so $\theta_1$ will drop. This will generate a drop in employment at routine occupations (“job polarization”). The effect on $\theta_0$ is unclear. Overall, we expect job polarization to be driven mostly by higher skilled workers. I will argue that this behavior closely fits the empirical findings on wage and employment patterns in the late 1980s and 1990s.

Wage polarization stops when low skilled workers start having a comparative advantage in routine occupations. This will occur when the first inequality in Equation 3.1 no longer holds for all $\theta_i$. At that point, higher skill routine workers at routine occupations will continue to leave. But some of the employment drop would be offset by joining into
Figure 3.3.: Change in log Wage for Small RBTC
routine occupations from the bottom of the skill distribution.

Finally, comparative advantage will be flipped, and inequality will start growing at the bottom. This will occur when

$$\forall \theta : \frac{\partial \log \varphi_R(\theta)}{\partial \theta} < \frac{\partial \log \varphi_M(\theta)}{\partial \theta}$$  \hspace{1cm} (3.2)

At this point, routine occupations employ the lowest skilled workers. Any further increase in $\tau$ will make wages relatively drop for the lowest paid workers as shown on Figure 3.4. A growing share of routine goods will be produced by capital. Hence, job polarization will continue as more workers will leave routine occupations. More routine goods will be produced, making workers in manual and abstract occupations more productive. This will increase their wages, and so increase inequality at all parts of the distribution. I will argue that this behavior fits empirical findings from mid 2000s onwards.

3.3. Skewness Decomposition

The main empirical tool I use in this paper to quantify RBTC in the data is skewness decomposition. Before diving into its details, I’ll briefly review alternative decomposition methods and why they are unable to capture the effects of RBTC based on the theoretical framework. I’ll then talk about how skewness decomposition can address all these challenges, and how it will be able to quantify the contribution of the predicted wage changes from the model.
Figure 3.4.: Changes in log Wages for Large $\tau$
3.3.1. Challenges to Standard Decomposition Methods

The first challenge with decomposing wage polarization is that it’s unclear which statistic we should decompose. When studying inequality, various indices such as variance of log wages, or Gini, could be used to measure inequality levels. These indices are useful to see how inequality varies across different time periods or countries. They can also be naturally decomposed, allowing researchers to better understand why inequality varies. Variance of log wages for instance, can be easily decomposed into a between and a within component

\[
V \left( \log w \right) = E \left[ V \left( \log w | X \right) \right] + V \left( E \left[ \log w | X \right] \right)
\]

This very simple decomposition allowed us to learn that a large share of the increase in inequality during the 1980s was due to an increase in inequality between education groups.\(^3\) Ideally, we would want to be able to perform a similar exercise for wage polarization.

However, there’s no clear single index to measure wage polarization. Since inequality is rising at the top but declining at the bottom, we cannot use the same indices to measure wage polarization. So far, we have used both the 90/50 and the 50/10 wage ratio to describe wage polarization. We could potentially use some combination of those two measures. However quantiles generally cannot be decomposed as elegantly, forcing us to use more general decomposition methods.

More general methods have their own drawbacks. Methods such as Juhn et al. (1993); DiNardo et al. (1996) are constructing counterfactual distributions in partial equilibrium, holding some components fixed. Since the full distribution is constructed, every statistic can then be calculated, making those methods very general. However, the interpretation

\(^3\)See Yitzhaki and Schechtman (2013) for Gini Decomposition.
of these methods is harder, as some arbitrary choices could potentially affect the results quite substantially. The order of components, which is usually completely arbitrary, has an effect on the contribution of each component. The choice of a baseline year is also arbitrary, and could affect the results as well (Lemieux, 2010). Moreover, those methods cannot accommodate any changing in the coding of the category that is used, making it hard to study long periods of time.

Most importantly, most methods don’t have the right component to capture the effect of RBTC. Most decomposition methods are decomposing the changes in the wage distribution into changes in composition, prices and an unexplained part. So for example, decomposing by occupations, will allow to study the effect of occupational transitions, and trends in mean wage at each occupation. As we discussed in Section 3.2.2, RBTC generates important trends within occupations. Inequality within manual, routine and abstract occupations could change substantially as well. Moreover, inequality could change in opposite directions in each occupation. This is why previous papers have not been able to show RBTC is generating most of wage polarization.

The closest approach to this paper is perhaps a re-centered influence function (RIF) regression. This method was used to study wage polarization in a paper by Firpo et al. (2013). This method doesn’t suffer from path-dependence, but still requires an arbitrary choice of baseline year. The main drawback of this method is that it doesn’t capture trends within occupations. Firpo et al. (2013) document that inequality trends within occupations are asymmetric, and inequality drops at routine occupations. But the RIF-regression cannot quantify the effect of these trends, as they are not reflected in occupation premiums (prices), nor in occupation compositions.

These methods are building counterfactual distributions allowing an additional component (such as composition, prices or residual) to vary each time. The effect of that component is the change in the statistic at each stage. But reordering the components will change the results (the effect of prices holding composition fixed or allowing it to change is different). See Fortin et al. (2011) for an extensive discussion on this.
Skewness decomposition will be able to address all these challenges.

3.3.2. Skewness Decomposition

Wage polarization can be measured with skewness. Skewness is the third standardized moment and is defined by

\[ S(Y) = E \left[ \left( \frac{Y - \mu}{\sigma} \right)^3 \right] \] (3.3)

It provides a measure for the asymmetry of the distribution relative to the mean. Figure 3.5 shows graphically the link to wage polarization by plotting the derivative of the empirical influence function at each quantile for a standard normal distribution. Intuitively, it shows the effect on skewness of a small increase in log wages for each quantile of the distribution, in case the log wage distribution was normally distributed. This shows that skewness increases exactly when wages at the edges increase compare to the middle. This pattern aligns quite well with the real trends in wages by quantile that I will show in Section 3.5.1.

The main advantage of using skewness is that it has a simple decomposition. Writing \( Y \) as standardized log wages, \( X \) some category we wish to decompose by, and \( \mu_3 \) as the third centralized moment we can write

\[ S(Y) = \mu_3(Y) = E[\mu_3(Y|X)] + \mu_3(E[Y|X]) + 3COV(E[Y|X], V[Y|X]) \] (3.4)

This decomposition was previously introduced by Mincer (1974), and it is the third moment analogous for the variance decomposition formula. Similar to variance decomposition, skewness decomposition break skewness into independent components as well. This means that there is no problem of path-dependence or any need to arbitrary define a baseline year.
Figure 3.5.: Derivative of EIF on Skewness for Standard Normal Distribution
Empirical influence function is a function from the value of a given observation $x_i$, to some statistic $T_n(x)$ (in this case, the empirical skewness), taking the other observations $x_{-i}$ as given. I calculate this for a sample of $n = 100$. I sample 1,000 samples of 100 observations from a standardized Normal distribution, and calculate numerically the derivative at the $k$th order statistic at the sample point. The figure shows the mean over the 1,000 samples of this derivative.
The first and second component are quite standard. $E[\mu_3(Y|X)]$ can be thought of as a “within” component. It captures the remaining skewness within each category. It should be high when our division into categories is orthogonal to the increase in skewness, and therefore can be thought of as a residual component. $\mu_3(E[Y|X])$ captures skewness between groups and captures skewness due to differences between group averages. This component will be high if the increase in wage polarization is because of changes in prices for all workers in a group. For instance, changes in occupation premiums, return to education etc.

The third component captures the correlation between the levels and inequality at each group. Formally, it is the covariance between the conditional mean and variance for each value of $X$. When highly paid groups also have larger inequality, inequality will be higher at the top than at the bottom, making the distribution more positively skewed. This component will allow to capture patterns like we predict from the theoretical framework, where higher paid occupations also have higher inequality. This pattern will be missed if we use a method that only decompose to occupational composition and prices. Therefore, it will be critical in order to test if wage polarization is indeed related to occupations and RBTC. This component will turn out to be capturing most of the increase in skewness during the period of wage polarization.

While in this paper I will use skewness decomposition to study wage polarization, it could be applied to any distribution where the third moment is of interest. There are various cases in economics where we know that the distribution is very skewed, and the level of skewness has important implications. Some examples are the distribution of the return to patents, firm productivity, the distribution of capital or raw wages (without logs). Any variation in these distribution across time or places can be analyzed with skewness decomposition. Similar decompositions exist for higher moments as well.\footnote{A general way to find the decomposition is to write $Y = E[Y|X] + \varepsilon$ and then use Newton’s Binomial}
3.4. Data

This paper uses two sources of data to get both, a large sample, and a panel structure.

The main analysis is done using the CPS outgoing rotation group. Since the theories I examine are related to the real price of labor, they are best captured using hourly wages. The CPS outgoing rotation group provides the most accurate representative sample of hourly wages (Lemieux, 2006). I use the years 1979-2012, that capture the early increase in inequality, wage polarization, and the return to increase in inequality. I use the same definition of the sample as (Acemoglu and Autor, 2011). Missing wages are dropped, which doesn’t seem to affect the results as I’ll show for the main results.

One important limitation of this data is its relatively high level of measurement errors. This problem is particularly severe at the edges of the distribution. Misreporting of working hours could lead to extremely high or extremely low values of hourly wages. As Figure 3.5 shows, skewness is especially sensitive to very high and very low incomes, making measurement error a large problem for this exercise. To deal with this, I drop the top and bottom 5% of the positive wages for the skewness analysis. The level of 5% was chosen in order to take the minimal cut, without substantial fluctuations between consecutive years in the skewness estimator. However, smaller cuts would also yield similar results, only noisier.

In most of the analysis I focus on the years between 1992-2002. The reason is that a significant change in occupational coding has taken place before and after this period so it

\[ E \left[ (E[Y|X] + \epsilon)^n \right]. \]

See also Appendix C.1.

I multiply the CPS weights in the number of hours worked to focus on the real price of an hour of labor as explained in Lemieux (2010), which is also consistent with the literature.

See Acemoglu and Autor (2011) for exact definition of sample sizes. I thank the authors for publicly sharing their cleaned data online.

Cornfeld and Danieli (2015) analyze skewness in the Israeli labor market, using the entire distribution since measurement errors are not as severe, and reach very similar conclusions.
is difficult to make informative comparisons to other years.\textsuperscript{9} As I will show, most of the increase in polarization occurred during this time period. I will separately implement several tests on the years before and after this period, and show that no similar trends occurred.

I maintain a consistent definition for routine occupations, similar to the previous literature. I first translate all occupational coding into the same coding, following Autor and Dorn (2013). I then define all administrative, operation and production jobs as routine, based on the 1-digit category. This is a similar classification to previous literature (for instance, Acemoglu and Autor, 2011) with one important exception - I do not classify sales occupations as routine occupations.\textsuperscript{10}

In order to analyze transition and selection into occupations, I also use panel data. Specifically, I use the Panel of Income Survey Data in Section (PSID) for the same years. This data was chosen due to its long panel. Because it only includes a small sample of workers, I do not use it for the main analysis. I use the full core sample (“SRC”), without weights. The over-sample of low income household and the immigrant samples that were added in the 1990s are not used.

\textsuperscript{9}Unlike other decomposition methods, skewness decomposition allows for the coding of $X$ to change. However we cannot rule out that some of the trends in each component is driven by changes in coding.

\textsuperscript{10}While information on task components suggests that sales is a routine occupation, I do not see the same wage nor employment patterns in these occupations. Including sales in routine occupations will not strongly affect the results as it is a small share of workers compared to the other routine occupations. However, it will make them weaker. This could be suggesting that sales occupations are not as easily automated as could be inferred from their O*NET description.
3.5. Results

3.5.1. Skewness Decomposition: Results

Using skewness decomposition, I will now show that wage polarization is indeed driven by occupational trends. This finding fits the theory that wage polarization is driven by RBTC as hypothesized by Autor et al. (2006). However, I’ll show that the effect is more nuanced than previously thought. The effect is not driven by a drop in wages at middle-skill occupation, but more due to the drop in inequality within low-paying occupations.

I will start by showing that skewness does indeed capture wage polarization well. Figure 3.6 shows the trend in skewness between 1979-2012. The rise in skewness aligns very well with the timing of wage polarization as depicted in Figure 3.1. Starting from the late 1980s, until the early 2000s, when the 90/50 gap is rising and 50/10 gap is falling, we see an increase in skewness.
The rise in skewness is driven by trends in all parts of the distribution. An increase in skewness occurs when the distribution becomes more tilted towards the left hand side. This means an increase in the gap between the middle and high wages and a decrease in the gap between the middle and low wages. Figure 3.7 plots a bin scatter of the change in wages between 1992-2002, for 20-quantiles. This generates a U-Shape that was previously shown by Autor et al. (2006, 2008). The U-Shape received qualitatively resembles the EIF derivative plotted in Figure 3.5. This implies that skewness has grown because of both the rise in wages at the top and at the bottom, making it a good fit to measure wage polarization.

I will decompose the rise in skewness of the distribution into three components, as de-
scribed in Equation 3.4 for different choices of grouping ($X$). I will focus on the period between 1992-2002 since other years have different occupational coding (see Section 3.4). As Figure 3.6 shows this time period includes a big portion of the increase in skewness. We can decompose by any variable that is in the data, such as occupations, industries, education etc. I will look for grouping where the increase in skewness is captured by the between and then covariance component. Since any increase in the within component could be thought of as part of the increase that is unrelated to this choice of $X$ variable. Decomposing by occupations can explain almost the entire rise in skewness. Figure 3.8 presents the decomposition by 3-digit occupational coding.\footnote{Appendix Figure C.1 performs the same exercise using imputed wages and reaches very similar results.} The figure draws the change in each component since 1992, and the sum of the three, which will always equal the total. The first interesting conclusion from this figure is the importance of occupations in explaining the trends in skewness. The within component, which captures the part that is unrelated to occupations is very small, and could also be the result of classification errors. Therefore, .089 of the .095 total rise in skewness (93%), can be attributed to occupations.

Most of the increase in skewness is due to the covariance component. The between component, which is driven by the trends in mean wages in each occupation, can explain only 15% of the overall trend. The majority (79%) of the increase is through the rising correlation between the mean and the variance of log wages in occupations. In other words, the growing correlation between wage levels and inequality levels at each occupation. As I discussed in Section 3.3.1, this type of correlation is not captured by other decomposition methods, which is why earlier work heavily underestimated the contribution of occupational trends.

The results are not driven by any other worker characteristic I can observe in the data. Since occupations are correlated with workers skills or industries it is important to verify that occupations are not just proxying for some other worker characteristics. In Appendix Figure C.1 performs the same exercise using imputed wages and reaches very similar results.
Figures 3.2 and 3.3 I show the same decomposition results by industry, and education and experience. Clearly, in those cases the within component is much larger, suggesting that great portions of the trend in skewness is unrelated to these categories. Moreover, I can show that most of the increase in the between and covariance component in those cases is due to their correlation with occupations. Appendix C.1 discusses how to decompose by more than one category using a linear model. Appendix Figures C.4, C.5 shows that by doing so we get that the increase is almost entirely through occupations.

These results strongly support the hypothesis that RBTC is generating wage polarization. The theory of RBTC argues that it affects differently workers performing different tasks. Occupations are the best proxy for tasks we have in most data sets. The fact that wage polarization, as measured with skewness, is driven by occupations greatly supports this explanation. Other possible explanations are not directly linked to occupations in any particular way.

However, these results also teach us new things about the way RBTC is affecting wages.
Earlier models of RBTC such as Autor et al. (2006); Acemoglu and Autor (2011), would have predicted that the effect will be captured in the “between” component. These models argue that there is a drop in the price of routine tasks, making wages fall equally for all workers in routine-heavy occupation. This would have been captured in the between component. The rise in the covariance component suggest that the effect is actually driven by the asymmetric trends in occupations, which I will now discuss.

3.5.2. Asymmetric Trends in Occupational Variances

The increase in correlation between wage levels and inequality that is driving wage polarization, could be driven by different explanations. It’s unclear whether this rise is because of trends in the wage levels, or wage inequality in occupations, or maybe occupational compositions. I will show that the main driver is the drop in inequality in at low-paying routine occupations. This is another support for the RBTC explanation for wage polarization.

During the 1990s the change in inequality within occupations was strongly correlated to the occupation wage level. High paying occupations saw an increase in inequality, and low paying occupations saw a decrease. Figure 3.9 shows this by plotting the change in variance of log wages from the beginning of the period (1992/3) to its end (2001/2) as a function of expected log wages. Changes in inequality, measured with the variance of log wages, are clearly correlated with the initial wage levels. This fact was also documented before by Lemieux (2007).

In fact, the trends in with-occupation inequality can explain the full rise in the covariance component. I use counterfactual partial-equilibrium wage distribution to show that. The
Figure 3.9.: Changes in Variance by Expectation of Low Wages for Top Occupations 1992/3-2001/2

Wages at the top and bottom 5% were dropped (see Section 3.4). Includes all occupations with at least 0.5% of the total working hours (top 47 out of 501 occupations that include 53% of the total working hours). The expected log wage is the average of the entire period (1992-2002), and the variance is the difference between the average of the first and last two years (I pool two years together to reduce errors due to small sample size). The line is the best linear fit to the points.

Source: CPS Outgoing Rotation Groups
covariance is calculated by

\[
COV (E [Y|X], V [Y|X]) = \sum_{x} Pr (X = x) E [Y|X = x] V (Y|X = x)
\]

Most of the increase stems from changes in the variance of log wages at different occupations \((V (Y|X = x))\). To show that, I will fix the share of workers and the expected log-wage in each occupation to their averages throughout the period. Thus, I will allow only the variance to vary between years. Formally, I will calculate a counterfactual partial-equilibrium covariance

\[
COV (E [\bar{Y}|\bar{X}], V [Y|X]) = \sum_{x} Pr (\bar{X} = x) E [\bar{Y}|\bar{X} = x] V (Y|X = x) \tag{3.5}
\]

for \(t\) between 1992 and 2002.

I find that the asymmetric trends in variance can explain the entire increase in covariance. Figure 3.10 compares the real value of the covariance to its counterfactual value from equation 3.5. The counterfactual trend closely follows the real trend. This means that if the share of workers and the mean log wage in each occupation were held fixed, we would still get the same increase in the covariance, and hence the same wage polarization. Letting the share of workers, or the expected log wage to vary while other factors are fixed does not yield any similar results. From this we infer that indeed the increase in covariance, and thus the increase in wage polarization is mostly the result of the asymmetric changes in occupational variances.

Wage polarization in the 1990s is driven by the drop in inequality at lower paid occupations. Figure 3.11 shows the trend in occupational inequality by decade. For each decade, it plots a bin scatter of the changes in variance of log wages, by occupation mean log wage decile. The increase in inequality at high paying occupations is a long-standing
Figure 3.10.: Covariance of Expectation and Variance of Log Wages at Different Occupations

Covariance of mean log wage and variance of log wage by occupation: $COV(E[\log w|occ], V(\log w|occ))$. Wages at the top and bottom 5% were dropped (see Section 3.4). Counterfactual covariance is calculated by fixing $E[\log w|occ]$, and the share of workers in each occupation to their average throughout the period (Equation 3.5.)

Source: CPS Outgoing Rotation Groups

trend. However, the 1990s are unique for their drop in inequality at low-paying occupations. This is why inequality is dropping at the bottom of the distribution only at the 1990s, generating wage polarization instead of an increase in wage inequality as in other decades.

The drop in inequality at low paying occupations is driven mostly by routine heavy occupations. Figure 3.12 plots the changes in variance for routine and non-routine occupations. I bin occupations by their initial income decile in 1992 separately for routine and non-routine occupations and plot the mean change in the variance of log wages between 1992-2002. While there’s some drop in inequality at low paying occupations that are non-routine, the trend is definitely stronger for routine occupations. This is in accordance with the findings in Firpo et al. (2013) who find using the O*NET data that routine occupations tend to have a stronger decrease in variance.

The most significant relative decline in wages, happened at the top of the income dis-
Figure 3.11: Change in $V[\ln w|occ]$ by $E[\ln w|occ]$ - Binned Scatter Plot

In each decade I bin occupations into 10 equal size bins of occupations, weighted by occupation size. Each point calculate the mean log wage at base year, and change in Variance from the start to the end of decade. In each decade, I de-mean the distribution of log wages, to have mean of zero. Each period includes the time period without changing to occupational coding. Hence, the three periods are 1979-1991, 1992-2002, 2003-2012.

Source: CPS Outgoing Rotation Groups
Figure 3.12: Change in \( V[\ln w|\text{occ}] \) by \( E[\ln w|\text{occ}] \) 1992-2002

I bin occupations separately for routine and non-routine occupations into 10 equal size bins (deciles) of occupations, weighted by occupation size. Each point calculate the mean log wage at base year, and change in Variance from the start to the end of decade 1992-2002.

Source: CPS Outgoing Rotation Groups
Figure 3.13: Change in Relative Wages 1992-2002, Routine and Non-Routine Occupations

I bin workers separately for routine and non-routine occupations into 10 equal size bins (deciles). Each point calculate the mean log wage at 1992, and change in log mean wage 1992-2002, for this decile.

Source: CPS Outgoing Rotation Groups

Figure 3.13 plots the (demeaned) wage trend in each wage decile, separately for workers in routine and non-routine occupations. Throughout the entire distribution, wages increased in non-routine occupations above the average increase in the labor market (that is pinned to 0). Wages relatively dropped in routine occupations, but not in an equal manner. The drop is much more significant at the top of the distribution. This generates a drop in inequality at routine occupations.

This decline could be driven by either a change in sorting into occupations, or real decrease in wages for higher skilled workers. In the next section I’ll use panel data to decide between these two explanations.
3.6. Decrease in Return to Skill - Evidence from Panel Data

The decline in top wages in routine occupations could be driven by different explanations, which can only be distinguished using panel data. It is well established that during the years of wage polarization, there is also job polarization - a significant drop in employment at routine occupations. If job polarization is changing the composition of routine occupations to have less workers from the top and the bottom of the skill distribution (as argued by Cortes, 2016), we will get exactly this trend of wage compression in routine occupations. In the theoretical framework, this is the case where \( \sigma = 1 \). It is also possible that demand decline is more significant for higher-skill workers, generating wage compression even for workers who don’t switch occupations. This is the case when new technology is substitutional to skill (\( \sigma > 1 \)). Using the PSID data I will show that the evidence support the latter explanation.

3.6.1. Decline in Routine Occupations

I’ll start by showing directly that both employment and wage premiums dropped in routine occupations, repeating some of the analysis of Cortes (2016).

The decline in employment is shown in Figure 3.14. The PSID uses the same occupational coding since 1980, allowing us to compare the share of workers in routine occupations consistently across various years. While in 1980, the share of routine workers in the PSID sample was close to half of all workers it dropped to about a third by 2011. The trend seems long and steady throughout the entire period, and not particularly stronger in any period of time.

I use a fixed effect model to study the trend in occupational wage premiums. I estimate the following standard fixed effect model
Figure 3.14: Share of Routine Workers - PSID
Share of routine workers from overall sample. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit.
Source: PSID

\[
\log w_{ijt} = \beta X_{it} + \lambda_{jt} + \theta_{i} + \varepsilon_{ijt}
\]  
(3.6)

where \(X_{it}\) includes a quadratic in experience, \(\lambda_{jt}\) is occupation by year fixed effect for three occupations: manual, routine and abstract and \(\theta_{i}\) is individual fixed effect.

I find the premium for routine occupations has decline steadily. Figure 3.15 shows the wage premium of routine workers compared to manual and abstract occupations \((\lambda_{R} - \lambda_{M}, \lambda_{R} - \lambda_{A})\). The premium for routine occupations seems to decline compared to both other alternatives.

By construction, this setting assumes occupational wage premiums are the same for all skill levels. To test that, we would need a more general model, that I will use in the next section.
Figure 3.15: Wage Premium for Routine Occupation Compared to Abstract/Manual

Wage premiums are estimated using a fixed effect model (Equation 3.6). Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Similarly, abstract includes all managerial, professional and technician occupations. Manual includes service, sales, and agriculture. Taking routine workers as the reference category, I plot minus one times the coefficient for manual and abstract by year.

Source: PSID
3.6.2. Interactive Fixed Effect Model

To test if returns to skills are changing over time, we need to go beyond the standard fixed effect model. In a fixed effect model, the log wage ratio of two workers with a skill level of $\theta_0, \theta_1$ will always remain $\theta_0 - \theta_1$, as long as they are in the same occupation. Wage compression within occupation is therefore impossible in this kind of model. Therefore, we would need a more general setting.

I will estimate an interactive fixed effect model, where returns to skill could change. Specifically, I will estimate the following equation

$$\log w_{ijt} = \beta_{ijt}X_{it} + \lambda_{jt} + \alpha_{jt}\theta_i + \varepsilon_{ijt}$$ (3.7)

The only difference from Equation 3.6 is that the individual fixed effects $\theta_i$ are interacted with an occupation time-varying coefficient $\alpha_{jt}$. This $\alpha$ parameter will be the focus of this analysis, as it measures the returns to skills in each occupation at each year. I will use either three categories of occupation, or 1-digit. I will estimate the model using a least square estimator (Bai, 2009).

Alternative ways to estimate this model yield very similar results. Least square method is consistent when the number of observations per individual is large enough. Since the estimates for $\hat{\theta}_i$ is estimated with noise, it could be correlated with $\varepsilon_{ijt}$. While this correlation is asymptotically zero, it’s unclear whether the number of periods used is sufficiently large. Therefore, I also estimate the model using an alternative approach.

I instrument for $\theta_i$ with worker’s years of education. Generally, other approaches must require an additional source of information, such as an instrument. For instance Holtz-Eakin et al. (1988) use lagged variables as instruments. Here, instead of assuming $\hat{\theta}_i$ and $\varepsilon_{ijt}$ are uncorrelated, I assume that years of education are uncorrelated

\footnote{For instance Holtz-Eakin et al. (1988) use lagged variables as instruments.}
with $\varepsilon_{ijt}$. This assumes that years of education is only affecting log wages through its effect on $\theta_i$.\(^{13}\)

### 3.6.3. Returns to Skill by Occupation

The estimation of Equation 3.7 supports the case of $\sigma > 1$ that is described in the model. That is, technological progress is substitutional to skills in routine occupations.

Supporting the assumption of the model, I find evidence that middle-skill workers have a comparative advantage in routine-intense occupations. Figure 3.16 plots the estimated density of $\theta_i$ for each one of the three occupational categories in 1981. At this point, routine occupations employ a large share of workers. On average, workers in manual occupations are the less skilled, workers in abstract occupations are the most skilled, and workers in routine occupations have an average level of skill. In the context of the model, since workers sort into occupations based on comparative advantage, this implies that returns to skill are highest at abstract occupation, then in routine, and smallest in manual. However, because manual occupations include a smaller share of all workers, the majority low skilled workers are still in routine occupations.

I test that using a one-dimensional skill is a reasonable assumption for this context. I do this by allowing $\theta_i$ to vary by $j$. This allows for a different skill to be used in each occupation, as in a Roy model. I find that the correlation between $\theta_{ij}$ for a given value of $i$ are between .66-.74, as shown in Table 3.16. Since we can only estimate the correlation for workers that chose to switch occupations, they might be upward biased. But more than half (51%) of workers do switch occupations and so most of the sample is used. Moreover, $\theta_{ij}$ is measured with a high level of noise, which biases the results downward. The high

\(^{13}\)Put differently, this approach assumes that trends in returns to education are identical to trends in returns to skill that are not captured with years of schooling. This is implied by the model that assumes only a one-dimensional skill.
Figure 3.16.: Estimated PDF of Skill (θi) by Occupation Category in 1981

Histogram for $\hat{\theta}_i$ from an interacted fixed effect model (Equation 3.7), by three occupation categories. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Similarly abstract includes all managerial, professional and technician occupations. Manual includes service, sales and agriculture.

Source: PSID
level of correlation suggest that we’re not losing much precision by allowing for only one skill.

The key finding from estimating the interactive fixed effect model is that there is a decline in returns to skill in routine occupations. The return to skill in routine occupation is measured with the parameter $\alpha_{Rt}$ in Equation 3.7. Since there is one degree of freedom in this equation, I pin $\alpha_{R,1980}$ to 1. At the beginning of the period, during the 1980s return to skill actually increase quite significantly. Then, exactly in the years of wage polarization and the decrease in inequality in routine occupations, there is a large drop in return to skill. This shows that during that time, wages compressed in routine occupations, even without any changes in their composition.

This pattern doesn’t repeat itself in every occupation. I estimate Equation 3.7 allowing the return to skills ($\alpha_{jt}$) to vary by 1-digit occupation and year. Figure 3.18 plots the coefficient for $\alpha_{jt}$ for each 1-digit occupation. Almost all occupations experienced an increase in return to skill during the 1980s. This fits the existing literature on the rise of inequality in the 1980, that argues that returns to education is sharply increasing in that period (Katz and Murphy, 1992).

Starting from the 1990s, the patterns are different by occupation. Administrative workers and Operators and to some extent also Production workers, the three occupations

<table>
<thead>
<tr>
<th></th>
<th>Abstract</th>
<th>Routine</th>
<th>Manual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Routine</td>
<td>.74</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Manual</td>
<td>.71</td>
<td>.66</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3.1: Correlation of Occupational Skills

Pearson correlation coefficient between $\tilde{\theta}_{ij0}$, $\tilde{\theta}_{ij1}$ for pairs of occupation categories. $\tilde{\theta}_{ij}$ are estimated using Equation 3.7, allowing $\theta_{ij}$ to vary by the three occupational categories. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Similarly abstract includes all managerial, professional and technician occupations. Manual includes service, sales and agriculture. Each correlation is calculated using all workers who worked at both categories.

Source: PSID
Figure 3.17: Returns to Skill ($a_{Rt}$) in Routine Occupations

Returns to skill are calculated in an interactive fixed effect model ($a_{Rt}$, using Equation 3.7). Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit.

Source: PSID
classified as routine, have a significant drop in return to skill. In contrast, in service occupations, return to skill have not shown any decline. This category includes most of the manual workers. During the 2000s it had a return value of 1.3, compared to 1.1-1.2 in administrative and operator occupations. This could potentially explain why workers from routine occupations decided to join service occupation (Autor and Dorn, 2013).

Abstract occupations vary in their trends in return to skills. For professional workers, it’s hard to see any decline in the return to skill, which is among the highest in any occupation. But managerial and technicians do seem to have some decline. This decline seems to be most prominent in the 2000s, which fits the theory of Beaudry et al. (2016) for a reverse in the demand trend for skilled labor. This requires further investigation that is beyond the scope of this paper.

As a result, decrease in the wage premium for routine occupations is sharper for higher skilled workers. In Figure 3.15 I followed previous literature that showed the decrease in routine premium for the average routine worker. But this masks significant heterogeneity by skill level. To account for those I plot the change in wage premiums accounting for the different levels in $\alpha_{jt}$, for skill levels at the 10th, 50th and 90th percentile.

Figure 3.19 shows the results comparing routine to abstract occupations. There is a significant decline in all parts of the skill distribution. This can explain why we see a large growth in the share of workers in abstract occupations. However, the decline is sharper for higher skilled workers. This implies that comparative advantage for skilled workers in abstract occupations became more significant.

The comparison of routine to manual occupations is even more striking. Figure 3.20 plots the result. During the 1980s, the premium for moving to routine occupations from manual occupations was positively correlated with skills. This generated a comparative advantage of higher skilled workers in routine occupations, compared to manual occupations.
Figure 3.18.: Return to Skill ($\alpha_{jt}$) by 1-digit Occupation

Return to skill are calculated in an interactive fixed effect model ($\alpha_{jt}$, using Equation 3.7). $\alpha_{jt}$ varies by 1-digit occupation and year.

Source: PSID
Figure 3.19.: Occupation Premium for Routine vs Abstract by Skill Percentile

Difference in log predicted wage for workers in routine versus abstract occupation for three percentiles at the distribution of $\theta_i$. $\theta_i$ are defined net of age and cohort. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Similarly abstract includes all managerial, professional and technician occupations.

Source: PSID
But during the 1990s the direction of the correlation flips. In the 2000s, the premium is higher for lower skilled workers, and is in fact even negative in some years for the higher skill ones. This implies that comparative advantage flipped during this period, making manual occupations more suitable for higher-skilled workers compared to routine. Overall, the decline in premium is close to zero for low skilled workers. This suggest that we can expect most of the drop in employment to be from higher skilled workers.

Appendix Figures C.6 repeats these results using years of education as instruments for skills. As discussed in the previous section, a least square approach would be inconsistent if the number of periods is not large enough. This is solved when instrumenting for $\theta_i$ with years of education. I find that the results are fairly similar, and the same conclusions
As predicted by the model and by the wage trends I estimate, job polarization is driven almost entirely from higher skilled workers. Figure 3.21 plots the decline in the share of routine workers separately for workers above and below the mean skill level. While in the early 1980s there has been an equal share of above and below mean workers in routine occupations, this is far from the case later. There has been a steady decline in the share of above mean workers in routine occupations, cutting their share in more than one half during this period. At the same time, the share of workers below mean in routine occupations stays fairly stable around 23% throughout most of the period. As a result, routine occupations gradually became low skill occupations, and are no longer middle-skill.

3.7. Discussion

This paper uses novel methods to address the main puzzles regarding the effect of RBTC on wages. The empirical findings closely align with the predictions of the model I outlined for such technological change. Together they explain why at first, in the 1990s wages relatively drop in the middle of the distribution, why later in the 2000s wages relatively drop at the bottom of the distribution, and why this was not captured by standard decomposition methods.

I show that the decrease in middle wages in the 1990s is the result of a technology that is replacing the usage of skills in routine occupations. I first use skewness decomposition to show that almost the entire trend of wage polarization is driven by occupational trends, which are impossible to capture with most other decomposition methods. I find that the 1990s are unique for their decrease in inequality within routine occupations. Even though
Figure 3.21.: Share of Routine Workers by Skill

Share of workers in routine occupations binned by level of $\theta_i$. $\hat{\theta}_i > 0$ is above mean, and otherwise below. $\theta_i$ are defined net of age and cohort. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Share is taken from all workers in sample.

Source: PSID
most routine workers are earning below median wages, most of the relative in decline in wages is around the middle, because of the wage compression in routine occupations.

Using an interactive fixed effect model I show that returns to skill declined in routine occupations. I use panel data to show that wages compressed in routine occupations even regardless of any employment trends. In the context of the model, this fits only the case where technology is substitutional to skills ($\sigma > 1$). Technology blurs the skill differences between workers, making all workers more equally productive. This generates the largest drop in demand for skilled workers in routine occupations. Supporting this notion, I find that job polarization is driven almost entirely by the decline in employment of higher-skill workers in routine occupations.

The model, and the empirical finding also provide an answer for why wage polarization stopped. Since the return to skills declined in routine occupations, middle skilled workers no longer have comparative advantage in them. The panel results show that gradually, routine occupations became low-skilled. As a result, starting approximately from early 2000s technological improvement in routine occupations do not affect middle wage workers anymore, since they no longer work there.

Other explanations do not fit these empirical patterns as well. Institutional changes, such as an increase in minimum wage could potentially generate a decline in middle wages by increasing lower wages. Decline in unions could also affect middle-wage workers more (Lemieux, 2007). High growth rates and low unemployment could also potentially boost lower wages, and so generate a relative decline in middle wages. However, neither of these explanations is expected to work through occupations in particular, more than through education levels or industries. The effect on occupations also should not be different for routine occupations, than it is on manual occupations such as service jobs.

An increase in demand for service occupations seems to be more of an outcome than
a cause to wage polarization. A demand increase in service occupations should have attracted more workers from the bottom of the skill-distribution. However, I find that job-polarization is mostly the result of employment drop for above average skill workers. The decrease in return to skill also fits the findings in research that study the causal effect of RBTC on firm wage distribution. Gaggl and Wright (2017) exploit natural experiments where exposure to technology varies by firms. They find that the new technology is generating wage compression within routine workers in a given firm. In this paper I quantify that this wage compression is actually the main driver of wage polarization.

Looking forward, if RBTC continues, its negative effect on low-income workers should only increase. As routine occupations have a larger share of low-skilled workers, any further shift towards capital in performing their main tasks will be more similar to a classic skilled-biased technological change. However, new technologies may start affecting the returns to skill in manual and abstract occupations. This will generate different wage trends beyond the scope of the theoretical framework I discussed here.

The results of this paper highlight the need for additional research on the micro-foundations of the decrease in return to skill in routine occupations. In the model, I used a general macro-level production function to explain how new technology replaces skill. This production function could be a result of a technology that is replacing the usage of human memory, or performs complicated calculations, making these skills less valuable. Hence, leaving workers with very basic simple tasks that don’t depend on skill as they don’t leave room for mistakes.

An alternative explanation is that the price of switching to the new technology is only justified when it saves the costs of the more skilled workers. In this case, the technology could in theory replace every worker. But its price will only be justified if workers are earning above its cost, which is why it’s replacing mostly the high earners. Using matched
employer-employee data to test if employers that are paying higher wages are adopting more technology could be an interesting exercise. However such data exist mostly in European countries that do not always experience the same wage trends (Massari et al., 2013).
Bibliography


URL http://www.nber.org/papers/w3572
URL http://www.aeaweb.org/articles.php?doi=10.1257/aer.103.5.1553


URL http://dx.doi.org/10.1162/rest.90.2.300


URL http://www.jstor.org/stable/2171802


URL http://dx.doi.org/10.1086/694153


URL http://www.nber.org/papers/w19843


Dweck, C. S., 1999. Self-theories: Their role in motivation, personality, and development.


URL http://qje.oxfordjournals.org/content/early/2014/05/08/qje.qju011.short


URL http://pubs.aeaweb.org/doi/abs/10.1257/aer.104.8.2509


URL http://www.nber.org/papers/w3927


URL http://www.jstor.org/stable/3003636


URL http://www.jstor.org/stable/30034058

URL http://link.springer.com/10.1007/s00148-007-0169-0


OECD, 2018b. Part-time employment rate (indicator).


URL http://dx.doi.org/10.1111/j.1468-0262.2002.00441.x


Robinson, J., 1933. The Economics of Imperfect Competition. Palgrave Macmillan.


URL http://www.jstor.org/stable/2662082


URL http://link.springer.com/content/pdf/10.1007/b98971.pdf
A. Apendix to Chapter 1

A.1. Theoretical Appendix

A.1.1. Continuous Logit Distribution

We follow Dagsvik (1994) in defining the continuous logit that produces $\varepsilon_{i,z}$ and $\varepsilon_{j,z}$. In this section we define the distribution of $\varepsilon_{i,z}$ and the distribution of $\varepsilon_{j,x}$ is defined similarly.

Every worker $i \in \mathcal{I}$ draws $\varepsilon_{i,z}$ shocks from a Poisson process on $\mathcal{Z} \times \mathbb{R}$ with intensity

$$f(z) \, dz \times e^{-\epsilon} d\epsilon$$

This is different from the Poisson process used in Dupuy and Galichon (2014) as the density $f(z)$ also affects the intensity, which allows this distribution to be properly defined over a larger class of functions for $\tau(x,z)$, including a constant, or simple polynomials. Denoting by $P_i$ the infinite but countable points chosen in the process, every worker has a set

$$\{\varepsilon_{iz} = \alpha \varepsilon | (z, \epsilon) \in P_i\}$$

This process yields a distribution of $\varepsilon_{i,z}$ that has several similarities to finite extremum
value type-1 distribution. These similarities are all derived from one basic property of this point process.

**Proposition A.1.** Let \( g : \mathbb{Z} \rightarrow \mathbb{R} \) be a function that satisfies

\[
\int_{\mathbb{Z}} e^{g(z)} f(z) \, dz < \infty
\]

and let \( S \subseteq \mathbb{Z} \) be some Borel measurable subset. Define

\[
\psi_g^S = \max_{z \in S \cap P_\mathbb{Z}} \{ g(z) + \varepsilon_{i,z} \}
\]

Then

\[
\psi_g^S \sim EV_1 \left( \alpha \log \int_{S} \exp \frac{g(z)}{\alpha} f(z) \, dz, \alpha \right)
\]

and

\[
S_1 \cap S_2 = \emptyset \iff \psi_{g_1}^{S_1} \perp \psi_{g_2}^{S_2}
\]

**Proof.** This proposition stems from the fact that in a Poisson process, the amount of points chosen in two disjoint Borel measurable sets \( B_1, B_2 \) has an independent distribution \( N(B_i) \sim Poisson(\Lambda(B_i)) \) with

\[
\Lambda(B_i) = \int_{B_i} \lambda(x) \, dx
\]

Therefore, in our context the cumulative distribution function of \( \psi_g^S \) is

\[
P(\psi_g^S \leq x) = P\left( N\left( (S \times \mathbb{R}) \cap \{ g(z) + \alpha \varepsilon > x \} \right) = 0 \right)
\]
From the Poisson distribution this is

\[
\log P(\psi g < x) = -\Lambda \left( S \times \left\{ \epsilon > \frac{x-g(z)}{\alpha} \right\} \right)
= -\int_S \int_{\frac{x-g(z)}{\alpha}} f(z) e^{-\epsilon} d\epsilon dz \\
= -\int_S e^{\frac{x-g(z)}{\alpha}} f(z) dz \\
= -\exp \left[ -\frac{x}{\alpha} \log \int_S \exp \frac{g(z)}{\alpha} f(z) dz \right]
\]

which is exactly a cumulative distribution function of \( EV_1 \left( \alpha \log \int_S \exp \frac{g(z)}{\alpha} f(z) dz, \alpha \right) \).

Since every draw of points in a Poisson process is independent, \( S_1 \cap S_2 \iff \psi g_{S_1} \perp \psi g_{S_2} \).

This Proposition has several important implications for our context. It implies that even though \( \epsilon_{i,z} \) is not defined for every \( z \in Z \), it is defined infinitely often for every Borel measurable subset that includes \( z \), and the maximum for that set \( \psi g \) has an extreme-value type-1 distribution.

Since workers in equilibrium are getting a sum of a continuous function (which we mark by \( \omega(x,z) \)) and \( \epsilon_{i,z} \) (Lemma 1.2), then we get that the maximum value they receive also has an \( EV_1 \) distribution, for every Borel measurable set of jobs. Moreover, the probability density to choose a particular observables \( z_j \) is similar to the finite case and its exact value is

\[
f(z_j|i) = f \left( z_j = \arg \max_{z \in S_i} \{ g(z) + \epsilon_{i,z} \} \right) = \frac{\exp \left[ \frac{1}{\alpha} \omega(x,z_j) \right] f(z_j)}{\int_Z \exp \left[ \frac{1}{\alpha} \omega(x,z) \right] f(z) dz}
\]

Another link to the finite multinomial logit can be drawn if we divide \( Z \) into a finite number of disjoint sets \( Z = \bigcup_{i=1}^n S_i, S_i \cap S_j = \phi \). Then the value of the best job for worker \( i \) in each subset \( (\psi g) \) is \( EV_1 \) distributed. The choice of the best job characteristics \( z_{m(i)} \) would be made with a finite multinomial logit, over these \( n \) options. When we increase \( n \), the sets become smaller, and the choice becomes closer to an infinite options choice.
Note that in a standard multinomial logit, increasing the number of options to infinity will yield an infinite compensation, but this is not the case here. This is because when the number of options \( n \) grow, the mean measure of \( S_i \) decreases in a rate of \( \frac{1}{n} \). Therefore the location parameter of each one of the choices, decreases in a rate of \( \frac{1}{n} \) as well from the proposition.

### A.1.2. Proofs

#### A.1.2.1. Proof of Lemma 1.1

Part 1: We will start by formally defining the densities we are using. We will use \( \mathcal{I}^\phi, \mathcal{J}^\phi \) to mark the set of unmatched workers and jobs.

**Definition A.1.** Let \( f(i,j): \mathcal{I} \times \mathcal{J} \rightarrow \mathbb{R}_{\geq 0} \) be the density that satisfies for every Borel measurable subset of potential matches \( B \subseteq \mathcal{I} \times \mathcal{J} \)

\[
\int_B f(i,j) \, di \, dj = \frac{\mu(B \cap M)}{\mu(M) + \mu(\mathcal{I}^\phi) + \mu(\mathcal{J}^\phi)}
\]

where \( \mu \) is the measure function.

Intuitively, this is the joint density of observing \( i \) and \( j \) matched in equilibrium. Similarly we define a density over the probability of observing worker and job with specific characteristics matched in equilibrium.

**Definition A.2.** Let \( f(x,z): \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}_{\geq 0} \) be the density that satisfies

\[
f(x,z) = \int_{X_i=x} \int_{Z_j=z} f(i,j) \, di \, dj
\]

From these definitions we can derive the conditional distribution of a match for a given worker.
**Definition A.3.** Let $f^i_j$ be

$$f^i_j = f(j|i) = \frac{f(i,j)}{f(i)} = \frac{f(i,j)}{I^{-1}}$$

**Part 2:** Let $i, i' \in I$, $j, j' \in J$ with $X_i = X_i'$ and $Z_j = Z_j'$. From Assumption 1.1 $\tau_{ij}$ has the same distribution as $\tau_{i'j'}$, and therefore $f(i,j) = f(i',j')$.

Hence, from Definition A.2,

$$f(X_i, Z_j) = I f(X_i) J f(Z_j) f(i,j)$$

and from Definition A.3

$$f^i_j = \frac{f(X_i, Z_j)}{f(X_i) f(Z_j)} J^{-1}$$

and we normalized $J = 1$.

**A.1.2.2. Proof of Lemma 1.2**

Let $i, i' \in I$ with $X_i = X_i' = x_0$ and $j, j' \in J$ with $Z_j = Z_j' = z_0$, where $m(i) = j$ and $m(i') = j'$. The sum of compensation equals the total surplus, hence

$$\omega_{ij} + \pi_{ij} = \tau(x_0, z_0) + \varepsilon_{i,z_0} + \varepsilon_{j,x_0}$$

$$\omega_{i'j'} + \pi_{i'j'} = \tau(x_0, z_0) + \varepsilon_{i',z_0} + \varepsilon_{j',x_0}$$

For stability, it must be that

$$\omega_{ij} + \pi_{i'j'} \geq \tau(x_0, z_0) + \varepsilon_{i,z_0} + \varepsilon_{j',x_0}$$

$$\omega_{i'j'} + \pi_{ij} \geq \tau(x_0, z_0) + \varepsilon_{i',z_0} + \varepsilon_{j,x_0}$$

200
Note that the sum of the two weak-inequalities is equal to the sum of the two equalities, therefore they must hold with equality (otherwise, the sum should hold both as an equality and strong inequality). Hence, we can rewrite

$$\omega_{ij} - \omega_{ij'} = \varepsilon_{i, z_0} - \varepsilon_{i', z_0}$$

$$\pi_{ij} - \pi_{ij'} = \varepsilon_{j, x_0} - \varepsilon_{j', x_0}$$

In other words, compensation for workers and employers in matches with the same characteristics is constant up to their value of $\varepsilon$, so we can write

$$\omega_{ij} = \omega(x_0, z_0) + \varepsilon_{i, z_0}$$

$$\pi_{ij} = \pi(x_0, z_0) + \varepsilon_{j, x_0}$$

$$\omega(x_0, z_0) + \pi(x_0, z_0) = \tau(x_0, z_0)$$

We can also pin down the alternative offers

$$\omega_{ij'} = \tau_{ij'} - \pi_{ij'} = \tau(x_0, z_0) + \varepsilon_{i, z_0} + \varepsilon_{j', x_0} - \pi(x_0, z_0) - \varepsilon_{j', x_0} = \omega(x_0, z_0) + \varepsilon_{i, z_0} = \omega_{ij}$$

This implies that all employers with $Z_j = z_0$ who are matched with $X_{m^{-1}(j)} = x_0$ are willing to make the same offer. Therefore, both workers and employers are facing a continuous logit choice. Hence, we can link the values of $\omega(x_0, z_0)$ and $\pi(x_0, z_0)$ to their choice probabilities (see Appendix A.1.1):

$$f(x_0|z_0) = \frac{\exp \left[ \frac{1}{\alpha} \pi(x_0, z_0) \right] f(x_0)}{\int_X \exp \left[ \frac{1}{\alpha} \pi(x, z_0) \right] f(x) \, dx}$$

The denominator is the expected value $\pi_j$, which is a function of $Z_j = z_0$ so we can rewrite
it as $\pi (z_0)$. Taking logs we get

$$\alpha \log f (x_0 | z_0) = \pi (x_0, z_0) + \log f (x_0) - \pi (z_0)$$

and with Lemma 1.1

$$\pi (x_0, z_0) = \alpha \log f_j^i + \alpha \log J + \pi (z_0)$$

therefore

$$\omega_{ij} = \tau (x_0, z_0) - \pi (z_0) + \alpha \log f_j^i + \varepsilon_{i,z_0}$$

where $J$ was pinned to 1.

**A.1.2.3. Proof of Lemma 1.3**

First equality is by definition, and because the first best and second best options are equivalent. We showed that $\omega_{ij} = \omega (x_i, z_j) + \varepsilon_{i,z_j}$. The expected compensation of worker $i$ is

$$\omega (x_i) = E [\omega^* (x_i, z_j)] + E [\varepsilon_{i,z_j}^*]$$

From the continuous logit structure we know that (similar to the previous proof)

$$\omega (x_i, z_j) = \alpha \log f_j^i + \omega (x_i)$$

hence

$$\omega (x_i) = E [\alpha \log f_j^i + \omega (x_i)] + E [\varepsilon_{i,z_j}^*]$$

Therefore

$$E [\varepsilon_{i,z_j}^*] = - \alpha \int f_j^i \log f_j^i dj$$
Similarly for \( \varepsilon_{j,x_0} \) and combinedly:

\[
E \left[ \varepsilon_{i,z_j}^* + \varepsilon_{j,x_0}^* \right] = -\alpha \int f_j \log f_j^i \, dj
\]

### A.1.2.4. Proof of Theorem 1.2

Following the notations from the previous proofs. The \( \omega_{ij} \) offer can be written as

\[
\omega_{ij} = \omega(x_i, z_j) + \varepsilon_{i,z_j}
\]

and \( \varepsilon_{i,z_j} \) is unaffected by \( \lambda \) hence

\[
\frac{d\omega_{i,j}}{d\lambda_i} = \frac{d\omega(x_i,z_j)}{d\lambda_i}
\]

In the previous proofs we showed that

\[
\omega(x_i,z_j) = \alpha \log f_j^i + \omega(x_i)
\]

\[
\pi(x_i,z_j) = \alpha \log f_j^i + \pi(z_j)
\]

hence

\[
\omega(x_i,z_j) - \pi(x_i,z_j) = \omega(x_i) - \pi(z_j)
\]

Adding \( \tau(x_i,z_j) \) and dividing by 2:

\[
\omega(x_i,z_j) = \frac{1}{2} (\tau(x_i,z_j) - \pi(z_j) + \omega(x_i))
\]
\(\tau(x_i, z_j), \pi(z_j)\) don’t change by the definition of \(\lambda\). Hence the only effect is on \(\omega(x_i)\).

\[
\frac{d\omega_{i,j}}{d\lambda_i} = \frac{1}{2} \frac{d\omega(x_i)}{d\lambda_i}
\]

We get the value for \(\omega(x_i)\) from the decomposition in Equation 1.6. Since \(\tau(x_i, z_{j'}), \pi(z_{j'})\) remain constant the remaining effect is on the OOI.

\[
\frac{d\omega_{i,j}}{d\lambda_i} = \alpha dOOI_i \frac{d\lambda_i}{d\lambda_i}
\]

A.1.2.5. Proof of Theorem 1.2

This is similar to before, only that \(\varepsilon_{i,z_j}\) is allowed to change as well. Since \(E[\varepsilon_{i,z_j}] = \alpha OOI\) we get the effect from the previous lemma, in addition to the effect on the OOI.

\[
\frac{d\omega_{i,j}}{d\lambda_i} = 2\alpha dOOI_i \frac{d\lambda_i}{d\lambda_i}
\]

A.1.2.6. Alternative Definitions for \(\lambda\)

Assume workers and equally distributed across the real line (as in Section 1.2.5). Each worker is a 3-dimensional tuple \((l_i, y_i, c_i)\) and \(\tau_{ij}\) is defined as

\[
\tau_{ij} = y_i - c_i |l_i - l_j| + \varepsilon_{ij}
\]

Now workers log density is a triangular function, with its peak at \(l_i\) (Laplace). Hence,

\[
f^j_i = \frac{c_i}{2} \exp -c_i |l_i - l_j|
\]
The OOI is (shifting $l_i$ to 0)

$$\int_0^\infty c \exp^{-cl} \left( \log \frac{c}{2} - cl \right) dl = \log \frac{c}{2} - 1$$

The mean value $E[\tau(x, z)]$ is

$$\int_0^\infty c \exp^{-cl} (y_i - cl) = y_i - 1$$

Hence, setting the commuting cost $c$ only affects the OOI but not net productivity and can be served as $\lambda$. This will also work for more general settings, as long as worker and job locations are not correlated with locations.

Another example is to define $\lambda$ as the intensity of the Poisson process for the continuous logit process. Higher $\lambda$ will mean more options on average in every subset of jobs.

**A.1.3. $f(x, z)$ Estimation**

To estimate a logistic regression following Equation 1.10, we maximize the following likelihood

$$\max_\theta \sum_k \log P(y_k|x_k, z_k; \theta)$$

where $\theta$ are the parameters defined in this equation, including matrix $A$. We rewrite Equation 1.10 in a more general form. Note $p_k(\theta) = P(Y_k = 1|X = x_k, Z = z_k)$:

$$\log \frac{p_k(\theta)}{1 - p_k(\theta)} = \sum_{j=1}^K \beta_j h_j(x_k, z_k)$$

where $K$ is the number of moments $h_j$ we control for in this regression.
Then the $K$ FOC of this maximization converge asymptotically to

$$E \left[ p_k (\theta) h_j (x_k, z_k) \right] = E \left[ h (x_k, z_k) | y_k = 1 \right] s$$

where $s = P (Y = 1)$ is the share of real data (in our case $\frac{1}{2}$). Using $\frac{p_k (\theta)}{1 - p_k (\theta)} = \frac{f(x, z)}{f(x) f(z)} \frac{s}{1 - s}$ we can write

$$E \left[ \frac{f (x, z)}{s f (x, z) + (1 - s) f (x) f (z)} h_j (x, z) \right] = E \left[ h_j (x, z) | \text{real} \right]$$

The RHS is simply the moment of $h_j (x, z)$ in the real data. The LHS is the moment of $h_j (x, z)$ in the full data (real and simulated), weighted by the probability it is real.

If the model is correctly specified and the functional form assumption on $\frac{f(x, z)}{f(x) f(z)}$ is true, $\theta$ will be estimated consistently. This is because

$$E \left[ \frac{f (x, z)}{s f (x, z) + (1 - s) f (x) f (z)} h_j (x, z) \right] = \int \frac{f (x, z)}{s f (x, z) + (1 - s) f (x) f (z)} h_j (x, z) (s f (x, z) + (1 - s) f (x) f (z)) dxdz =$$

$$= \int h (x, z) f (x, z) dxdz = E \left[ h (x, z) | \text{real} \right]$$

If the model is misspecified, our estimate of $\frac{f(x, z)}{s f(x, z) + (1 - s) f(x) f(z)}$ will not be converging to the real density ratios. Instead we will equalize moments of some other weighted average of $h_j$

$$E \left[ w (x, z, \theta) h_j (x, z) \right] = E \left[ h_j (x, z) | \text{real} \right]$$

where

$$w (x_k, z_k, \theta) = s^{-1} \frac{\exp \sum_{j=1}^{K} \beta_j h_j (x_k, z_k)}{1 + \exp \sum_{j=1}^{K} \beta_j h_j (x_k, z_k)}$$

We next analyze these weights as $s \to 0$. We will mark $h_1 (x, z) = 1$, the offset of the
regression. When \( s \to 0 \), \( \frac{p(\theta)}{1-p(\theta)} \to 0 \) as well, therefore \( \exp \sum_{j=1}^{K} \beta_j h_j(x_k, z_k) \to 0 \). With some abuse of notation, we will redefine \( \beta_1 \) as \( \beta_1 - \log s \). Therefore

\[
\lim_{s \to 0} w(x, z, \theta) = \exp \sum_{j=1}^{K} \beta_j h_j(x_k, z_k) = \frac{f(x, z)}{f(x) f(z)}
\]

The density of the full data approaches the density of the simulated data. Hence overall, we get

\[
E[w(x, z, \theta) h_j(x, z) | \text{sim}] = E[h_j(x, z) | \text{real}]
\]

In order to calculate the OOI, we simulate values from \( f(x) f(z) \), and reweight them based on \( \frac{f(x, z)}{f(x) f(z)} \). This is because we hold workers fixed, and simulate \( z \) values from \( f(z) \). As \( s \to 0 \) we use weights that converge to \( w(x, z, \theta) \). The above equation guarantees that we sample from a distribution with same moment value for every \( h_j(x, z) \), even if the model is misspecified.

Dupuy and Galichon (2014) produce a distribution with the same second moments as the data, and same marginal distributions. Therefore, when \( s \to 0 \), and \( h \) include all \( X, Z \) interactions, and an indicator for every \( x_k \), and every \( z_k \) value (that is, \( h(x, z) = 1_{x=x_k} \) or \( h(x, z) = 1_{z=z_k} \) for every \( k \)), we get the same distribution.

### A.1.4. Standard Errors for a Wald Estimator with Matching

We want to estimate the standard errors of \( \hat{\alpha} \), defined in Equation 1.17. Both the nominator (reduced form), and the denominator (first stage) are standard matching estimators for average treatment effect on treated (ATET). Abadie and Imbens (2006) show how to estimate standard errors for ATET. But to estimate correctly the standard error for the Wald estimator, we also need to estimate the covariance of the first stage and reduced form. So we extend their approach for this case.
Mark the ATET on log wages (reduced form), and OOI (first stage) as:

\[
\rho = E[\log w (1) - \log w (0) | T = 1]
\]

\[
\gamma = E[OOI (1) - OOI (0) | T = 1]
\]

where (1) means value when treated and (0) when not treated. \( T \) is treatment status (so this is the mean effect for treated).

The Wald estimator is then

\[
\alpha = \frac{\rho}{\gamma}
\]

For each match \( m \) (treated unit and one or more control unit), define

\[
X_m = \left( \begin{array}{c}
X_{1m} \\
X_{2m}
\end{array} \right) = \left( \begin{array}{c}
\log w (1) - \log w (0) \\
OOI (1) - OOI (0)
\end{array} \right)
\]

Our estimators are then simply

\[
\left( \begin{array}{c}
\hat{\rho} \\
\hat{\gamma}
\end{array} \right) = \overline{X_m}
\]

Asymptotically

\[
\overline{X_m} \sim N\left( \begin{array}{c}
\rho \\
\gamma
\end{array} , \begin{pmatrix}
\sigma_{11} & \sigma_{12} \\
\sigma_{12} & \sigma_{22}
\end{pmatrix} \right)
\]

With the Delta method

\[
V(\hat{\alpha}) = \frac{1}{\gamma^2} \left( \sigma_{11} - 2 \frac{\rho}{\gamma} \sigma_{12} + \frac{\rho^2}{\gamma^2} \sigma_{22} \right)
\]

Abadie and Imbens (2006) tells us how to find \( \sigma_{11}, \sigma_{22} \) which are \( V(\hat{\rho}), V(\hat{\gamma}) \). We want to extend their approach to \( \sigma_{12} \).
The challenge in getting the variance correctly for matching with replacement, is that the matches are not independent. Some observations from the control pool appear in more than one match. Following Abadie Imbens we write

\[
V (\bar{X}_m) = \frac{1}{N_1} \sum_m (X_m - \bar{X}_m)^T (X_m - \bar{X}_m) + \frac{1}{N_1} \sum_{T=0} (K_i (K_i - 1)) \hat{V}_i
\]

with

\[
\hat{V}_i = \hat{V} \begin{pmatrix} \log w_i \\ OOI_i \end{pmatrix}
\]

where \( N_1 \) is number of treated units, \( K_i \) is the number of times observation \( i \) from the control pool was used. \( \hat{V}_i \) is a 2x2 matrix of the variance for that particular observation. The first part is a standard variance calculation. The second part corrects for the covariance between the matches.

If an observation \( i \) is used \( K_i > 1 \) times, then there are \( K_i (K_i - 1) > 0 \) pairs of matches that both use it, and so their covariance is not 0, but includes \( \hat{V}_i \).

To estimate \( \hat{V}_i \) we follow Abadie and Imbens (2006) and use nearest neighbor from the control group. So for every control observation we find a match from the control group as well and write

\[
\hat{V}_i = \frac{1}{2} \begin{pmatrix} \log w_i - \log w_{m(i)} \\ OOI_i - OOI_{m(i)} \end{pmatrix}^T \begin{pmatrix} \log w_i - \log w_{m(i)} \\ OOI_i - OOI_{m(i)} \end{pmatrix}
\]

This is asymptotically unbiased.

In practice, the only difference from Abadie and Imbens (2006) is that we also have a covariance component.

\[
COV (\hat{\rho}, \hat{\gamma})
\]
Which we estimate with

\[
\frac{1}{N_1} \sum \left( \log w_m(1) - \log w_m(0) \right) \left( OOI_m(1) - OOI_m(0) \right) - \hat{\rho} \hat{\gamma} +
\]

\[
2 \cdot \frac{1}{2} \cdot \sum_{T=0}^{K_i} K_i (K_i - 1) \left( \log w_i - \log w_m(i) \right) \left( OOI_i - OOI_m(i) \right)
\]

If our two variables were the same (log \( w = OOI \)) then this would be the standard Abadie and Imbens (2006) formula for variance, as expected.

A.2. Data Appendix

A.2.1. LIAB

In this section we clarify the coding of some of the variables we use. Our panel data, allows us to observe some variables several times in the data, and correct for coding errors. In particular, we set German citizenship to one, if this worker was ever reported as a German citizen by her employer.

We also take the highest level of education we observed until every year. All upper secondary school certificates are coded as upper-secondary. In some years intermediate and lower secondary education are coded with the same value. In these cases, if we observe the worker in other years and can infer their schooling level we use that. Otherwise, we code these workers in a separate category for either lower or intermedia secondary education.

For training occupation, we use the occupation in which workers spent the longest time in training. The LIAB data specify whether a worker is in vocational training and their occupation. For the large majority of workers, there is only one occupation in which they
perform their vocational training. In rare cases where workers have conducted training in more than one occupation, we use the occupation in which the training was longer. If the we never observe the worker during vocational training, we take the occupation in which they conducted an internship. If this is unobserved as well, we use the first occupation they were observed in, as long as at least ten years have passed since we first observed them.

We calculate distance at the district level. For each district, we calculate the district center, by taking the weighted average of the latitude and longitude coordination of each city in this district. We then calculate the distance between the districts, taking into account the concavity of the earth.

A.2.2. BIBB Survey

In this section we describe in more detail the BIBB survey and PCA analysis.

We use data from the 2011-2012 wave of the German Qualification and Career Survey conducted by the Federal Institute of Vocational Training (BIBB) and the Institute for Labor Market Research (IAB). The data cover 20,000 employed individuals between the ages of 16 and 65. We run PCA on this survey by questions category and aggregate the results by 2-digit industry and 3-digit occupations. We link the results to our main data. The top question in each category are shown in Table 1.3.
B. Appendix to Chapter 2

B.1. Pilot Survey Implementation Guide

B.1.1. Project Timeline

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 28, 2017 - April 30, 2017</td>
<td>Phone Survey</td>
</tr>
<tr>
<td>April 6, 2017 - August 2, 2017</td>
<td>Address Based Survey</td>
</tr>
</tbody>
</table>

B.1.2. Sample Design

B.1.2.1. Population

The target population for the study consisted of non-institutionalized persons age 18 and over living in Los Angeles County, California (hereafter: L.A. County).

B.1.2.2. Sampling Frame

Numbers for the landline sample were drawn by Survey Sampling International with equal probabilities from active blocks (area code + exchange + two-digit block number) that contained one or more residential directory listings. The cellular sample was drawn
by Survey Sampling International through a systematic sampling from 1000 blocks dedicated to cellular service according to the Telcordia database. The sampling frame excluded nontelephone households. The address-based sample was drawn by Marketing Systems Group using a stratified cluster design. Census tracts were used as primary sampling units (PSU) with a systematic probability proportionate to size selection. Within each of the 110 sampled PSUs, 55 addresses were selected for a total of 5,500 addresses from which 3 self-representative replicates were created. The sampling frame excluded vacant, seasonal, educational, P.O. boxes not flagged as the only address where the owner receives mail, and drops (multi-residence dwellings with no unit number). The strata are described below.

Stratification

The ABS sample stratified the universe of Census tracts in L.A. County by poverty rates. We obtained tract-level poverty estimates from the American Community Survey 2015 5-year Summary File, using 100% of the Federal poverty line as our benchmark. County-level poverty rates were then Z-scored. Tracts with poverty rates more than one standard deviation below the mean were assigned to the Low Poverty (hereafter, “Low”) stratum, tracts with poverty rates more than one standard deviation above the mean were assigned to the High Poverty (hereafter, “High”) stratum, and tracts with poverty rates between -1 and 1 standard deviations were assigned to the Medium Poverty (“Medium”) stratum. Tracts for which no information on poverty is available (those with very few residents) were assigned to the Medium stratum.

Respondent Selection For the telephone and CAPI modes, interviewers asked to speak with an adult age 18 or older, living in the household. The web invitations and paper surveys were sent via USPS with the assumption that adults were opening and reading the households mail.
Screening and Eligibility The survey screened for respondents age 18 or older who live in L.A. County. Anyone providing a ZIP Code outside of L.A. County was screened out and were not eligible for the survey.

B.1.3. Comparison Between Address Based Sampling and Phone Based Sampling

<table>
<thead>
<tr>
<th></th>
<th>Address-Based</th>
<th>Phone-Based</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(1) = (2)</td>
</tr>
<tr>
<td>Age</td>
<td>47.406</td>
<td>49.157</td>
<td>0.220</td>
</tr>
<tr>
<td>Female</td>
<td>0.606</td>
<td>0.557</td>
<td>0.202</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.425</td>
<td>0.444</td>
<td>0.633</td>
</tr>
<tr>
<td>Black</td>
<td>0.094</td>
<td>0.129</td>
<td>0.160</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.358</td>
<td>0.373</td>
<td>0.694</td>
</tr>
<tr>
<td>Asian</td>
<td>0.141</td>
<td>0.064</td>
<td>0.002</td>
</tr>
<tr>
<td>Other race</td>
<td>0.059</td>
<td>0.041</td>
<td>0.302</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than HS</td>
<td>0.050</td>
<td>0.064</td>
<td>0.453</td>
</tr>
<tr>
<td>Some HS, Incomplete</td>
<td>0.056</td>
<td>0.074</td>
<td>0.357</td>
</tr>
<tr>
<td>High School/GED/Incomplete College</td>
<td>0.294</td>
<td>0.332</td>
<td>0.301</td>
</tr>
<tr>
<td>Two Year Associate Degree</td>
<td>0.094</td>
<td>0.077</td>
<td>0.448</td>
</tr>
<tr>
<td></td>
<td>Address-Based</td>
<td>Phone-Based</td>
<td>p-value</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>--------------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Bachelor or Graduate Degree</td>
<td>0.506</td>
<td>0.453</td>
<td>0.183</td>
</tr>
<tr>
<td><strong>Household Income</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 15,000</td>
<td>0.146</td>
<td>0.189</td>
<td>0.157</td>
</tr>
<tr>
<td>15,000 to less than 25,000</td>
<td>0.109</td>
<td>0.148</td>
<td>0.164</td>
</tr>
<tr>
<td>25,000 to less than 35,000</td>
<td>0.097</td>
<td>0.110</td>
<td>0.617</td>
</tr>
<tr>
<td>35,000 to less than 50,000</td>
<td>0.128</td>
<td>0.076</td>
<td>0.040</td>
</tr>
<tr>
<td>50,000 to less than 75,000</td>
<td>0.152</td>
<td>0.121</td>
<td>0.282</td>
</tr>
<tr>
<td>75,000 to less than 100,000</td>
<td>0.125</td>
<td>0.095</td>
<td>0.250</td>
</tr>
<tr>
<td>100,000 or more</td>
<td>0.243</td>
<td>0.261</td>
<td>0.612</td>
</tr>
<tr>
<td>Observations</td>
<td>300</td>
<td>343</td>
<td></td>
</tr>
</tbody>
</table>
B.2. Survey Implementation Guide

B.2.1. Project Timeline

<table>
<thead>
<tr>
<th>Date Range</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 5, 2017 - April 9, 2018</td>
<td>Phone screens conducted</td>
</tr>
<tr>
<td>October 30, 2017</td>
<td>Mailed out study brochures</td>
</tr>
<tr>
<td>November 20, 2017 - May 7, 2018</td>
<td>In-person interviews conducted</td>
</tr>
<tr>
<td>May 21, 2018</td>
<td>Final dataset received</td>
</tr>
</tbody>
</table>

B.2.2. Sample Design

B.2.3. Population

The target population for the study consisted of adults age 18 and over who self-reported having grown up poor and currently living in Shelby County, TN, Tulsa County, OK, or Jefferson and Orleans Parishes, LA.

B.2.4. Sample

**Sampling Frame**  Numbers for the landline sample were drawn by Survey Sampling International (SSI) with equal probabilities from active blocks (area code + exchange + two-digit block number) that contained one or more residential directory listings. The cellular sample was drawn by Survey Sampling International through a systematic sampling from 1000 blocks dedicated to cellular service according to Telcordia, an FCC approved national telephone database administrator. The sampling frame excluded non-telephone households. Each sampling frame is described in detail below.
**Landline sample**  When creating its landline RDD database, SSI starts with a computer file of over 53 million directory-listed households. Using area code and exchange data regularly obtained from Telcordia and additional databases, this file is subjected to an extensive cleaning and validation process to ensure that all exchanges are currently valid, assigned to the correct area code, and fall within an appropriate set of ZIP Codes. Telephone exchanges and 100 blocks (i.e., the last two digits of the telephone number) that contain one or more listed residential telephone numbers are considered valid and are represented in the database. The RDD database is formed of all telephone numbers having such valid exchanges and working blocks.

Each exchange is assigned to a single county. For those exchanges that cross county and/or state lines, the exchanges are assigned in their entirety to the county with the highest number of listed phones within that exchange. Abt Associates selects random digit samples from the RDD database using the Random A procedure. Random A is an SSI term denoting a systematic sample of random digit telephone numbers selected with equal probability across all working blocks. Within a county, the sampling interval is calculated by dividing the number of working blocks by the number of sampling points requested. Abt Associates uses a working block value of one to minimize under-coverage.

**Cell Phone Sample**  Mobile samples are selected from a database that contains all possible numbers in 100-blocks dedicated to wireless service and 100-blocks providing shared services but that have no directory-listed telephone numbers. SSI selects EPSEM (equal probability of selection methodology) samples selected from the cell phone number database. Blocks are in ascending order by exchange and block number within exchange, within county. Once the quota has been allotted to all the counties in the frame, a sampling interval is calculated for each county by summing all the working blocks in the county.
and dividing that sum by the number of sampling points assigned to the county. From a random start between zero and the sampling interval, blocks are systematically selected from each county. Once a block has been selected, a two-digit random number in the range 00-99 is appended to the block to form a ten digit telephone number.

In order to more efficiently reach cell phone respondents, the cell sample was appended with activity code information by SSI. These activity codes provided information about the working status of each number in the cell phone sample. In the cell sample, 67% of numbers were flagged as “active” and 33% were flagged as “inactive”. Completely excluding all cell phone numbers classified as “inactive” from dialing could result in some coverage bias if some of those numbers were actually active, as observed in the pilot study. For this reason, cell phone numbers flagged as “inactive” were sub-sampled at approximately 50% rate.

**Billing ZIP Code-Matched Cell RDD Sample** Using billing ZIP Code-matched cell RDD sample can help to address the low geographic eligibility rate observed in the pilot, reducing screening costs. ZIP-matching is, however, only possible for cell phones with billing records. (The precise details of ZIP-matching methodology are proprietary and held closely by SSI and its competitors.) Therefore, SSI appended billing ZIP Code to every cell phone selected in the sample. From that we observed that 30.6% of the numbers were matched to a billing ZIP code inside the target areas, 11.8% were matched to a billing ZIP code outside the target areas and 57.6% of the numbers were not matched to a billing ZIP code. In order to improve the efficiency of the sample, all phone numbers matched to a billing ZIP code outside the targeted areas were excluded from the sample and phones numbers that were not matched to any billing ZIP code were sub-sampled at approximately 65%. All phone numbers matched to a billing ZIP code inside the targeted
areas were eligible to be dialed.

**Respondent Selection**  For the landline sample, interviewers asked to speak with an adult age 18 or older, living in the household. For the cell sample, interviews were conducted with the person who answered the phone and was age 18 or older.

**Screening and Eligibility**  The telephone screener first confirmed the age of the respondent and the zip code where the individual lived. If the respondent was 18 or over and lived in one of the eligible geographic areas, the survey then also screened for respondents who self-identified as having “grown up poor”. Persons who reported not having grown up poor were not eligible for the in-person extended interview, but their basic demographic information, such as gender, age, education and race/ethnicity, were collected for weighting purposes.
B.2.5. Number of Respondents by Market

<table>
<thead>
<tr>
<th></th>
<th>Shelby</th>
<th>Tulsa</th>
<th>Jefferson</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td>Total Phone Screens</td>
<td>2281</td>
<td>2237</td>
<td>1941</td>
<td>6459</td>
</tr>
<tr>
<td>Ineligible from Phone Screens</td>
<td>1334</td>
<td>1336</td>
<td>1172</td>
<td>3842</td>
</tr>
<tr>
<td>Eligible but Declined</td>
<td>526</td>
<td>489</td>
<td>375</td>
<td>1390</td>
</tr>
<tr>
<td>Eligible and Agreed</td>
<td>421</td>
<td>412</td>
<td>394</td>
<td>1227</td>
</tr>
<tr>
<td>In-Person Interviews</td>
<td>308</td>
<td>308</td>
<td>312</td>
<td>928</td>
</tr>
<tr>
<td>Removed Well off Respondents</td>
<td>300</td>
<td>301</td>
<td>299</td>
<td>900</td>
</tr>
</tbody>
</table>

B.3. Data Description and Coding of Variables

B.3.1. Survey Variables and Indices

B.3.1.1. Outcome Variables:

1. Income:
   a) Household Income – This outcome was collected from the survey question, “What was your total annual household income last year, in 2016, before taxes?” When respondent received benefits from the government, responses were collected from the question, “What was your total annual household income last year, in 2016, before taxes, as you reported to the government?” If the respondent did not provide an answer to the above two questions, they were asked to
give the range in which their household income existed

- Under $10,000
- $10,000 to $20,000
- $20,000 to $30,000
- $30,000 to $40,000
- $40,000 to $50,000
- $50,000 to $75,000
- $75,000 to $100,000
- $100,000 to $150,000
- More than $150,000

If the range was specified, then household income for different ranges were taken as

- $10,000
- $15,000
- $25,000
- $35,000
- $45,000
- $62,500
- $87,500
- $125,000
- $150,000
b) Adjusted Household Income – Adjusted household income was calculated using the formula used for the Census Supplement of Poverty Measure.

i. For a household without children

\[ N = (adults)^5 \]

ii. Single parent household

\[ N = (1 + 0.8 \times \text{first child} + 0.5 \times \text{other children})^7 \]

iii. All other

\[ N = (adults + 0.5 \times \text{children})^7 \]

Then adjusted income is

\[ \frac{HH - Income}{N} \]

c) Individual Income – This outcome was collected from the survey question, “If you are not the only income earner in your household, what was your individual annual income last year, in 2016, before taxes?” When respondent received benefits from the government, responses were collected from the question, “If you are not the only income earner in your household, what was your individual annual income last year, in 2016, before taxes, as you reported to the government?” If the respondent did not provide an answer to the above two questions, they were asked to give the range in which their individual income existed and were recoded in the same way as household income above.

2. Adult Physical Health: This outcome was coded from the question, “In general, how is your health: excellent, very good, good, fair or poor?” where 1 = poor, 2 =
fair, 3 = good, 4 = very good, and 5 = excellent.

3. Adult Drug and Alcohol Use: This outcome was calculated as the total of 4 questions –

- Have you ever felt that you ought to cut down on your drinking or drug use?
- Have people annoyed you by criticizing your drinking or drug use?
- Have you ever felt bad or guilty about your drinking or drug use?
- Have you ever had a drink or used drugs first thing in the morning to steady your nerves or get rid of a hangover?

We coded each question to be 1 if they responded “yes” and if their earliest episode happened after the age of 18; or if they responded “yes” and their most recent episode happened after the age of 18. Variables were coded as 0 if they responded “no”, or if they responded “yes” but their earliest and most recent episode happened before the age of 18 years.

4. Adult Mental Illness: This outcome is calculated as the total of standardized responses to the following question –

For each of these, please indicate how often you have been bothered by it over the last 2 weeks.

- Feeling nervous, anxious or on edge
- Not being able to stop or control worrying
- Trouble relaxing
- Being so restless that it is hard to sit still
- Becoming easily annoyed or irritable
- Feeling afraid as if something awful might happen
• Little interest or pleasure in doing things
• Feeling down, depressed or hopeless
• Trouble falling asleep, staying asleep or sleeping too much
• Feeling tired or having little energy
• Poor appetite or overeating
• Feeling bad about yourself or that you are a failure or have let yourself or your family down
• Trouble concentrating on things, such as reading the newspaper or watching television
• Moving or speaking so slowly that other people could have noticed. Or, the opposite being so fidgety or restless that you have been moving around a lot more than usual
• Thoughts that you would be better off dead or of hurting yourself in some way.

where each statement’s response is coded as 1 = not at all, 2 = several days, 3 = more than half of days, 4 = nearly every day.

B.3.1.2. Demographics:

1. Gender - Coded from phone screener as “male” or “female”.

2. Race - Coded from phone screener as “black”, “hispanic”, “white” or “other”.

3. Education - Coded from phone screener as
   • No high school
   • Incomplete high school
• High School/GED/Incomplete College
• Two year Associate Degree
• Bachelor or some Post-Graduate Degree

4. Parental Income -

• Would you say your family during that time was pretty well off financially, about average, or poor?

• While you were growing up, from birth until age 16, did financial difficulties ever cause you or your family to move to a different place?

• From birth until age 16, was there a time when you or your family received help from relatives because of financial difficulties?

• From birth until age 16, was there a time of several months or more when your father had no job?

• Did your mother ever get Aid to Families with Dependent Children or welfare from your birth until age 16?

• How often do you remember the following happening to your parents or guardian until you were age 16? Would you say always, often, sometimes, rarely, or never?
  
  – Being unable to find child care or being forced to take a child out of child care because they couldn’t pay?
  
  – Falling behind in rent or mortgage payments?
  
  – Falling behind in gas, electric, or phone bills?
  
  – Being unable to pay for adequate transportation to get to work or school?
  
  – Being unable to get medical care because of the cost?
– Having trouble paying a credit card balance?
– Having too little money to buy enough food?
– Being a victim of a crime?
– Having a problem with alcohol or drug abuse?

B.3.1.3. Childhood Experience:

1. Lifetime Trauma Before 18
   • Before you were 18 years old, did you ever have to do a year of school over again?
   • Before you were 18 years old, were you ever in trouble with the police?
   • Before you were 18 years old, did either of your parents drink or use drugs so often that it caused problems in the family?
   • Before you were 18 years old, were you ever physically abused by either of your parents?

2. Any Lifetime Trauma - Please indicate whether the event occurred at any point in your life
   • Have you ever been homeless?
   • Have you ever been in a major fire, flood, earthquake, or other natural disaster?
   • Did you ever have a life-threatening illness or accident?
   • Did your spouse or a child of yours ever have a life-threatening illness or accident?
   • Have you ever fired a weapon in combat or been fired upon in combat?
• Has your spouse, partner, or child ever been addicted to drugs or alcohol?

• Were you the victim of a serious physical attack or assault in your life?

• Has a child of yours ever died?

3. Adverse Childhood Experience -

• Did a parent or other adult in the household often swear at you, insult you, put you down, or humiliate you or act in a way that made you afraid that you might be physically hurt?

• Did a parent or adult in the household often push, grab, slap, or throw something at you or ever hit you so hard that you had marks or were injured?

• Did an adult or person at least 5 years older than you ever touch or fondle you or have you touch their body in a sexual way or try to actually have oral, anal, or vaginal sex with you when you were a minor? (IF NEEDED: A minor is someone under the age of 18.)

• Did you often feel that no one in your family loved you or thought you were important or special or your family didn’t look out for each other, feel close to each other, or support each other?

• Did you often feel that you didn’t have enough to eat, had to wear dirty clothes, and had no one to protect you or your parents were too drunk or high to take care of you or take you to the doctor if you needed it?

• Were your parents ever separated or divorced?

• Was your mother or stepmother often pushed, grabbed, slapped, or had something thrown at her or sometimes or often kicked, bitten, hit with a fist, or hit with something hard or ever repeatedly hit over at least a few minutes or threatened with a gun or knife?
• Did you live with anyone who was a problem drinker or alcoholic or who used street drugs?

• Was a household member depressed or mentally ill or did a household member attempt suicide?

• Did a household member go to prison?

4. Risky Attitudes as Teenager -

• I had trouble concentrating or paying attention

• I lied or cheated

• I teased others a lot

• I disobeyed my parents

• I had trouble sitting still

• I had a hot temper

• I would rather have been alone than with others

• I hung around with kids who got into trouble

• I disobeyed at school

• I didn’t get along with other kids

• I had trouble getting along with teachers

• First, I would like to ask you about smoking habits. As a teenager, did you smoke cigarettes?

• As a teenager, did you ever have a drink of an alcoholic beverage? By a drink we mean a can or bottle of beer, a glass of wine, a mixed drink, or a shot of liquor. Do not include childhood sips that you might have had from an older person’s drink.
• As a teenager, did you ever use marijuana (that is grass or pot)?

• Excluding marijuana and alcohol, as a teenager, did you ever use any other drugs like cocaine or crack or heroin, or any other substance not prescribed for you by a doctor, in order to get high or to achieve an altered state?

5. Beliefs About Success

• When you were young, did you believe you would grow up to be successful?

• Was there a time in your life when help could have made all the difference?

B.3.1.4. Health Indices:

1. Mental Illness Before 16 - Before 16 did you have

• Depression

• Drug or alcohol problems

• Any other emotional or psychological problems

2. Physical Illness Before 16

3. Resilience -

• I tend to bounce back quickly after hard times

• I have a hard time making it through stressful event

• It does not take me long to recover from a stressful event

• It is hard for me to snap back when something bad happens

• I usually come through difficult times with little trouble

• I tend to take a long time to get over set-backs in my life

4. Locus of Control
• “What happens to me is my own doing” or “Sometimes I feel that I don’t have enough control over the direction my life is taking.”

• “When I make plans, I am almost certain that I can make them work” or “It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.”

• “In my case getting what I want has little or nothing to do with luck” or “Many times we might just as well decide what to do by flipping a coin.”

• “Many times I feel that I have little influence over the things that happen to me” or “It is impossible for me to believe that chance or luck plays an important role in my life.”

5. Growth Mindset

• You have a certain amount of intelligence, and you really can’t do much to change it.

• Your intelligence is something about you that you can’t change very much.

• You can learn new things, but you can’t really change your basic intelligence.

6. Grit

• New ideas and projects sometimes distract me from previous ones.

• Setbacks don’t discourage me.

• I have been obsessed with a certain idea or project for a short time but later lost interest.

• I am a hard worker.

• I often set a goal but later choose to pursue a different one.

• I have difficulty maintaining my focus on projects that take more than a few months to complete.
• I finish whatever I begin.
• I am diligent.

7. Self Esteem

• On the whole, I am satisfied with myself.
• At times, I think I am no good at all.
• I feel that I have a number of good qualities.
• I am able to do things as well as most other people.
• I feel I do not have much to be proud of.
• I certainly feel useless at times.
• I feel that I’m a person of worth, at least on an equal plane with others.
• I wish I could have more respect for myself.
• All in all, I am inclined to feel that I am a failure.
• I take a positive attitude toward myself.

8. Self Control

• I am good at resisting temptation.
• I have a hard time breaking bad habits.
• I am lazy.
• I say inappropriate things
• I do certain things that are bad for me, if they are fun.
• I refuse things that are bad for me.
• I wish I had more self-discipline.
• People would say that I have iron self-discipline.

• Pleasure and fun sometimes keep me from getting work done.

• I have trouble concentrating.

• I am able to work effectively toward long-term goals.

• Sometimes I can’t stop myself from doing something, even if I know it is wrong.

• I often act without thinking through all the alternatives.

9. IPIP

• I am the life of the party.

• I feel little concern for others.

• I am always prepared.

• I get stressed out easily.

• I have a rich vocabulary.

• I don’t talk a lot.

• I am interested in people.

• I leave my belongings around.

• I am relaxed most of the time.

• I have difficulty understanding abstract ideas.

• I feel comfortable around people.

• I insult people.

• I pay attention to details.

• I worry about things.
• I have a vivid imagination.
• I keep in the background.
• I sympathize with others’ feelings.
• I make a mess of things.
• I seldom feel blue.
• I am not interested in abstract ideas.
• I start conversations.
• I am not interested in other people’s problems.
• I get chores done right away.
• I am easily disturbed.
• I have excellent ideas.
• I have little to say.
• I have a soft heart.
• I often forget to put things back in their proper place.
• I get upset easily.
• I do not have a good imagination.
• I talk to a lot of different people at parties.
• I am not really interested in others.
• I like order.
• I change my mood a lot.
• I am quick to understand things.
• I don’t like to draw attention to myself.
• I take time out for others.
• I ignore my duties.
• I have frequent mood swings.
• I use difficult words.
• I don’t mind being the center of attention.
• I feel others’ emotions.
• I follow a schedule.
• I get irritated easily.
• I spend time reflecting on things.
• I am quiet around strangers.
• I make people feel at ease.
• I am exacting in my work.
• I often feel blue.
• I am full of ideas.

10. Diet

B.3.1.5. Family Indices:

1. Family Environment
2. Family Network
3. Relationship with Parents
4. Parenting
• How often would your parent say that you did something that gave him/her pleasure and enjoyment?

• How often would your parent say that you did something that greatly irritated him/her and got on his/her nerves?

• How often did your parent read to you?

• How often did your parent physically punish you as a child, for example by a spanking?

• How often did your parent praise you as a child, by saying something like “Good for you!” “What a nice thing you did!” “Thank you!” or “That’s good going!”

• How often did your parent tell you about his/her experience, by saying something like, “I saw a pretty bird outside just a little while ago”, or “I exercised so hard that I got really tired”, or “I was able to give some directions today to somebody that got lost”, or “I really like the way the sky looks now”.

• How often did you and your parent talk or play with each other, focusing attention on each other for five minutes or more, without your parent asking or telling you to do anything?

• How often did your parent tell you to do something, with an irritated or angry tone of voice?

• How often did you and your parent engage in make-believe play, where you each played the part of a character, and together made up a story to act out with each other?

• How often did you laugh with your parent?

• How often did your parent yell or speak to you in a very loud voice, with
irritated or angry emotion?

5. Adults you Could Trust 100%

6. Number of Adult Relationships Trusted

7. Quality of Adult Relationships

8. Adult Living with at Age 9

B.3.1.6. Neighborhood Indices:

1. Probability of Bottom 25 in Top 20 percentile

2. Fraction with Fathers Present

3. Neighborhood Safety Index
   - How safe were the parking lots and sidewalks near your neighborhood school?
   - How safe did you feel at home alone at night?
   - How safe were the streets near your home during the day?
   - How safe were the streets near your home at night?
   - Was anyone’s purse, wallet, or jewelry snatched from them?
   - Was anyone threatened with a knife or gun?
   - Was anyone beaten or assaulted?
   - Was anyone stabbed or shot?
   - Did anyone try to break into your home?
B.3.2. NLSY

The National Longitudinal Survey of Youth 1979 (NLSY79) is a nationally representative sample of 12,686 young men and women who were between 14 and 22 years old when they were first interviewed in 1979. These individuals were surveyed annually through 1994 and biannually thereafter. Covered topics include family background and demographic characteristics; household composition; schooling and aptitude information; income and assets; health conditions; alcohol and substance use; attitudes and aspirations; and more. For our analysis, we defined growing up poor to be an individual who we observe in the 2014 sample and in 1979 had a Poverty Status equal to one, an NLSY created indicator variable based on number of respondents in the household, family income and yearly poverty income guidelines established by the U.S. Department of Health and Human Services. For current income, we used the individual’s total income in 2014. Our interview used the same four questions as the NLSY79 to measure Rotter’s Locus of Control Scale. For additional information on the NLSY79 see the NLS Handbook. (https://www.bls.gov/nls/handbook/2005/nlshc2.pdf)

B.3.3. Dataset from Dobbie and Fryer 2013

The dataset was collected for Dobbie and Fryer, 2013 and is constructed from two sources: school-specific data collected from principal, teacher, and student surveys, lesson plans, and videotaped observations of classroom lessons; and administrative data on student demographics and outcomes from the New York City Department of Education (NYC-DOE). In 2010, the authors attempted to collect data from all charter schools in New York City with students in grades 3-8. Eligible schools were invited to participate via e-mail and phone and offered a stipend. Of the 62 eligible charter elementary schools (entry grades of PK to fourth) and 37 eligible charter middle schools (entry grades of fifth to
eighth), 26 elementary schools, and 13 middle schools chose to participate in the study. Within the set of participating schools, 19 elementary schools, and 10 middle schools were also able to provide usable admissions lottery data. A wide variety of information was collected from participating schools including details on teacher and staff development, instructional time, data driven instruction, and parent outreach obtained through a principal interview. Information on curricular rigor was coded from lesson plans collected for each testable grade level in both math and ELA. Finally, information on school culture and practices was gathered during full day visits to each school. Within each domain, an indicator variable was coded to equal to one if a school has an above median level of that input, selecting the variable or combination of variables that best captures the variation described by the qualitative literature. The administrative data from the NYCDOE included information on student race, gender, free and reduced-price lunch eligibility, behavior, attendance, and state math and ELA test scores for students in grades 3-8. Additional details are available in the Data Appendix of Dobbie and Fryer 2013.

**B.4. Technical Proofs**

Proof of proposition 1 is in the text.

Proof of Proposition 2

Write

$$\mu_1 = \mu_0 + \nu$$

For simplicity assume that $\mu_0 = 0$, hence $X$ is demeaned. We assume

$$f (X) = X \beta$$
And from OLS estimation
\[ \hat{\beta} = \left( X^T X \right)^{-1} X^T Y \]

Hence
\[ \Omega (I) = \nu^T \left( X^T X \right)^{-1} X^T Y \]

The Kullback-Liebler divergence of two multivariate distributions is
\[ D_{KL} (P (X|I = 1) \| P (X|I = 0)) = \frac{1}{2} \left( \nu^T \Sigma_0^{-1} \nu + tr \left( \Sigma_0^{-1} \Sigma_1 \right) + \ln |\Sigma_0| - \ln |\Sigma_1| - k \right) \]

where \( k \) is the number of variables. Taking our constants, and replacing \( \Sigma_0 \) with its estimator \( \left( X^T X \right) \) this is
\[ \frac{1}{2} \left( \nu^T \left( X^T X \right)^{-1} \nu + tr \left( \left( X^T X \right)^{-1} \Sigma_1 \right) - \ln |\Sigma_1| \right) \]

Together, we maximize
\[ \max_{\nu, \Sigma_1} \nu^T \left( X^T X \right)^{-1} X^T Y - \frac{\lambda}{2} \left( \nu^T \left( X^T X \right)^{-1} \nu + tr \left( \left( X^T X \right)^{-1} \Sigma_1 \right) - \ln |\Sigma_1| \right) \]

The first thing to notice is that \( \nu \) and \( \Sigma_1 \) are not interacting in this expression. Therefore this can be written as two separate exercises.

The solution \( \Sigma_1 = \Sigma_0 \) is from the fact that \( D_{KL} \) is minimized when the distributions are equal, and \( \Sigma_1 \) doesn’t effect \( \Omega (I) \).

To get \( \nu \) we solve
\[ \max_{\nu} \nu^T \left( X^T X \right)^{-1} X^T Y - \frac{\lambda}{2} \left( \nu^T \left( X^T X \right)^{-1} \nu \right) \]

FOC is
\[ \left( X^T X \right)^{-1} X^T Y - \lambda \left( X^T X \right)^{-1} \nu = 0 \]
\[ \nu = \frac{1}{\lambda} X^T Y = COV(X, Y) \]

Proof of Proposition 3-4:

We will prove a stronger version of both propositions such that Proposition 3 and 4 would be specific cases. We will add weights and controls.

Theorem: Assume

\[ P_i (X = x|I = 0) = \begin{cases} 
\omega(x) \exp \left( -\frac{\text{dist}(x, x_{i'})^2}{2\sigma^2} \right) & \exists x_{i'} \in \text{data s.t. } x_{i'} = x \\
0 & \forall x_{i'} \in \text{data s.t. } x_{i'} = x
\end{cases} \]

are distributed as in Proposition 4 and Assumptions 1 & 2 hold.
B.5. Phone Screen Instrument
Hello, my name is _____ and I’m calling from Abt Associates, a survey research firm on behalf of EdLabs at Harvard University (IF NEEDED: That is, the Education Innovation Laboratory). We’re conducting a brief research survey about how people’s background and upbringing affects their lives and future resources. It is called the “Understanding Inequality Survey.” It will only take 10 minutes of your time. LANDLINE ONLY: May I please speak to an adult, age 18 or older? (IF NEW RESPONDENT, RE-INTRODUCE AND CONTINUE) CELL ONLY: Are you 18 years old or older? (if No, screen out)

Thank you, as I mentioned the survey is about your upbringing and how you are doing today. We’ll start with some questions about your background and then move on to questions about your current situation.

All of your responses will be kept confidential and the reported results of the study will combine the responses of yours and others. No personal identifying information will be reported. Participation is voluntary and you can stop at any time. Some of the questions may be considered sensitive in nature and you may skip any question you prefer not to answer. This call may be monitored for quality assurance. Would you like the contact information for the researchers or the Committee on the Use of Human Subjects in Research at Harvard University, who approved this study?

[IF YES: If you have any questions about the survey, you can call 877-699-4340 or email InequalitySurvey@abtassoc.com. Additionally, this research has been reviewed by the Committee on the Use of Human Subjects in Research at Harvard University. They can be reached at 617-496-2847, 1350 Massachusetts Avenue, 9th Floor, Suite 935 Cambridge, MA 02138 and cuhs@harvard.edu.]

(If cell phone: If you are now driving a car or doing any activity requiring your full attention, I need to call you back later.) If you do not have any questions, may we begin?

1. To begin, for classification purposes only, can you please tell me your age as of today? RANGE 18-98, 99=Don’t know/Refused
IF Q1=99 ASK Q1a; ELSE code into appropriate category

1a. Are you between the ages of:

1. 18-24
2. 25-34
3. 35-44
4. 45-54
5. 55-64
6. 65 or more
7. (VOL) Don’t know/Refused

2. To make sure all areas are represented, can you please tell me which zip code you live in?

1. [DISPLAY APPROPRIATE ZIP CODES BASED ON MARKET]
98. (VOL) Other (specify) – Screen out
99. (VOL) Don’t know/Refused – Screen out

3. And how long have you lived in the [READ IN RESPONSE from Q2] zip code? Would you say… [read list]

1. Less than one year
2. 1-2 years
3. 3-4 years
4. 5-9 years
5. 10-14 years
6. 15-20 years
7. More than 20 years?
8. (VOL) My whole life
9. (VOL) Don’t know/Refused

ASK IF Q3=1 or 2, ELSE SKIP TO Q4

3a. What zip code did you live in before moving to [READ IN RESPONSE from Q2]?

[NUMERIC ENTRY, VERIFY MATCHES ZIP CODE FORMAT]
IF Q3=8 OR (Q1a=1 AND Q3=7) AUTOPUNCH Q4=1 AND SKIP TO Q5; ELSE, READ:

4. Where did you spend the majority of your time living as a child, between the ages of 5 and 15? Was it… (READ LIST)
   1. The neighborhood you currently live in
   2. Another neighborhood in [Shelby County]/[Tulsa County]/[your Parish]
   3. Somewhere else in [Tennessee/Oklahoma/Louisiana]
   4. A different state
   5. Overseas
   9. (VOL) Don’t know/Refused

5. How many people currently live in your household, including yourself? Please also include any children, under the age of 18.
   1. RANGE: 1-15
   99. (VOL) Don’t know/Refused

ASK IF Q5>1

5a. Is there anyone living in your household who is NOT in your immediate family. If so, how many? This includes people such as extended family (cousins, grandparents), friends, and roommates.
   (IF NEEDED: By immediate family, we mean a spouse/domestic partner, any siblings, parents, or children.)
   1. RANGE 1-number from Q5
   99. (VOL) Don’t know/Refused

6. How would you rate your current financial situation, would you say it’s… (READ LIST)
   1. Excellent
   2. Good
   3. Only Fair, or
   4. Poor
   9. (VOL) Don’t know/Refused

7. Have you, or have any immediate family members in your household ever received welfare or public assistance benefits? This includes SNAP (“snap”) or food stamps, TANF (“Tan-Eff”), or other housing benefits.
   1. Yes
   2. No
   9. (VOL) Don’t know/Refused
ASK IF Q7=1 (Yes)

7a. Are you or are any immediate family members in your household receiving such benefits now?
   1. Yes
   2. No
   9. (VOL) Don’t know/Refused

8. Do you consider yourself to have grown up poor?
   1. Yes
   2. No
   9. (VOL) Don’t know/Refused

9. Do you consider yourself poor now?
   1. Yes
   2. No
   9. (VOL) Don’t know/Refused

Thank you. Now I have just a few final questions to help us classify your answers.

10. I have to verify, what is your gender?
    1. Male
    2. Female
    3. (VOL) Other
    9. (VOL) Don’t know/Refused

11. What is the highest level of education you have received?
    1. Less than high school (Grades 1-8 or no formal schooling)
    2. High school incomplete (Grades 9-11 or Grade 12 with NO diploma)
    3. High school graduate (Grade 12 with diploma or GED certificate)
    4. Some college, no degree (includes some community college)
    5. Two year Associate’s degree from a college or university or community college
    6. Four year college or university degree/Bachelor’s degree (e.g., BS, BA, AB)
    7. Some postgraduate or professional schooling, no postgraduate degree (e.g., some graduate school)
    8. Postgraduate or professional degree, including master’s, doctorate, medical or law degree (e.g., MA, MS, PhD, MD, JD, graduate school)
    9. (VOL) Don't know/Refused

245
12. What is your race? (Accept multiple responses)
   1. White/Caucasian
   2. Black/African American
   3. Hispanic
   4. Asian/Asian-American
   5. Some Other Race (Specify)
   6. (VOL) Native American/American Indian/Alaska Native
   7. (VOL) Pacific Islander/Hawaiian
   9. (VOL) Don’t know/Refused

ASK Q13a OF LL SAMPLE ONLY

13a. Now thinking about your telephone use…does anyone in your household, including yourself, have a working cell phone?
   1. Yes
   2. No
   8. (VOL) Don’t know
   9. (VOL) Refused

ASK Q13b OF CELL SAMPLE ONLY

13b. Now thinking about your telephone use…is there at least one telephone inside your home that is currently working and is not a cell phone?
   1. Yes
   2. No
   8. (VOL) Don’t know
   9. (VOL) Refused

IF Q8>1, SKIP TO QCLASS

ASK Q14 of CELL SAMPLE ONLY

14. Those are all of the questions I have. For speaking with us today, we’d like to mail you a check for $5. All I’ll need is your name and address. Please note, this will be used for your check only and will not be associated with your survey responses.

   Name: 
   Address: 
   City: 
   State: 
   Zip: 

246
15. Thank you very much for answering my questions. Based on your answers to this survey, you qualify for the next round of our study which consists of an in-person interview. If you complete the full interview, you will be paid $150 as a thank you for your time. [IF NECESSARY: The $150 will be in the form of a pre-loaded Visa or MasterCard gift card.] The survey will take about two hours to complete and many people find the questions very interesting. An interviewer will come to interview you. [IF NEEDED: The interviewer will visit your home.] Can I take your contact information so that the interviewer can contact you to make a time for the interview?

1. Agrees to participate - CONTINUE
9. Refused – SKIP TO QCLASS.

16. Great. To start, I’ll just need your name and address.

[USE STANDARD ADDRESS BLOCK, FOR CELL VERSION, READ-IN INFO FROM Q14, IF REFUSED, SKIP TO QCLASS]

17. And can I please have the best phone number to reach you?

1. Enter phone [CONTINUE]
9. Refused [SKIP TO Q19]

17a. Is this a landline, cell, or work number?

1. Landline
2. Cell
3. Work

18. Is there another phone number that we can try if we can’t reach you there?

1. Enter phone [CONTINUE]
9. Refused [SKIP TO Q19]

18a. Is this a landline, cell, or work number?

1. Landline
2. Cell
3. Work

IF Q17a=2 OR Q18a=2, ASK Q21, ELSE – SKIP

19. Do we have your permission to contact you via text message to your cell phone?

1. Yes
2. No
20. Can I please have your email address?
   1. Provided email
   2. Don’t have email
   9. Refused

21. Sometimes when we try to contact people, we aren’t able to reach them because they may have moved or got a new telephone number. I’d like to get the contact information for 2 other people who don’t live with you that would know how to reach you? Let’s start with the first person.

   [USE STANDARD NAME/ADDRESS TEMPLATE]

21a. What is their relationship to you?
   1. Parent
   2. Brother/Sister
   3. Son/Daughter
   4. Other Family Member
   5. Friend
   6. Neighbor
   7. Other (Specify:______)
   9. (VOL) Refused

21b. And what is their phone number?
   1. Enter phone
   9. Refused

22. Can I have the second person’s name and address?

   [USE STANDARD NAME/ADDRESS TEMPLATE]

22a. What is their relationship to you?
   1. Parent
   2. Brother/Sister
   3. Son/Daughter
   4. Other Family Member
   5. Friend
   6. Neighbor
   7. Other (Specify:______)
   9. (VOL) Refused
22b. And what is their phone number?

1. Enter phone
9. Refused

Thanks so much. That is the end of the survey. In the next few weeks you’ll be getting a letter from us with more details about the in-person interview. An interviewer will also call you to schedule the interview at a time that’s convenient. If you have any questions about the survey you can reach out to the project directors at 877-699-4340, or email InequalitySurvey@abtassoc.com.

Would you like the contact information for the researchers or the Committee on the Use of Human Subjects in Research at Harvard University, who approved this study?

[IF YES: If you have any questions about the survey, you can call 877-699-4340 or email InequalitySurvey@abtassoc.com]. Additionally, this research has been reviewed by the Committee on the Use of Human Subjects in Research at Harvard University. They can be reached at 617-496-2847, 1350 Massachusetts Avenue, 9th Floor, Suite 935, Cambridge, MA 02138 and cuhs@harvard.edu.]

QCLASS: (DUMMY PUNCH, DO NOT ASK, USE THIS FOR QUOTA)

1. Eligible, agrees to next interview [COMPLETES FULL INTERVIEW]
2. Eligible, refuses next interview [REFUSAL AT Q15 OR Q16]
3. Not Eligible [Q8>1]

Thank you very much for your assistance. Have a nice day!
B.6. In-person Interview Instrument
Hello, my name is [ ]. May I please speak with _____?

Thank you for taking the time to talk with me today. I work for Abt Associates. Abt Associates is an independent research company and we are helping Harvard EdLabs with its Understanding Inequality Survey. We are conducting interviews with people who agreed to be in this study. You might remember completing a telephone interview on [TELEPHONE INTERVIEW DATE].

This interview will include questions on your background, behaviors, childhood experiences, attitudes and personality. It will take about 2 hours to complete. When we are done, you will be paid a $150 Visa gift card, as a way of saying thank you for your time.

Your participation in this study will help us understand how to develop policies that assist in social mobility and better understand the characteristics of people living in your neighborhood.

Before we begin the survey, I would like to assure you that all of your responses on this survey will be kept private; your name will not appear in any written reports we produce. Your responses to these questions are completely voluntary. That means you may choose not to answer any question, or you may stop the interview if you wish, but we hope you don’t. Some of the questions could be triggering to individuals with a history of post-traumatic stress disorder, or survivors of abuse and violence. You may find these questions upsetting. Please feel free to ask to skip these questions or discontinue the interview without penalty. You will not need to explain why you do not wish to answer or stop the interview. We will also provide you with a list of resources to local support organizations if you wish.

The information you provide will be kept private and only used for this study. By participating in this study, you will be greatly helping us further our understanding of the reasons some people do better in life and others do not.

Would you like the contact information for the researchers or the Committee on the Use of Human Subjects in Research at Harvard University, who approved this study?

[IF YES: If you have any questions about the survey, you can call 1-877-699-4340 or email InequalitySurvey@abtassoc.com. Additionally, this research has been reviewed by the Committee on the Use of Human Subjects in Research at Harvard University. They can be reached at 617-496-2847, 1350 Massachusetts Avenue, 9th Floor, Suite 935, Cambridge, MA 02138, or cuhs@harvard.edu.]

Do you have any questions before we begin?

USE STANDARD ROC/DISPOS
Before we get started, I just want to give you this copy of your rights as a participant and I also have a list of resources that may be useful.

[Interviewer: Provide Informed Consent and Resources List to Respondent]

Now, if you don’t have any questions, we can get started.

Adult Characteristics

Auto-fill race, gender, etc. from screener

Base: All

Q1 What is your religious preference: is it Protestant, Catholic, Jewish, some other religion or do you have no preference?

Do not read responses, code all that apply

[Programmer: Accept multiple responses]

1 Protestant (Go to Q1a)
2 Catholic (Go to Q2)
3 Jewish (Go to Q2)
4 Mormon (Go to Q2)
5 Greek Orthodox (Go to Q2)
6 Russian Orthodox (Go to Q2)
7 Muslim (Go to Q2)
8 Buddhist (Go to Q2)
9 Hindu (Go to Q2)
10 Jehovah’s Witness (Go to Q2)
11 Atheist (Go to Q2)
12 Agnostic (Go to Q2)
13 Something else (Go to Q2)
14 No preference (Go to Q2)
15 Nothing in particular (Go to Q2)
96 No other mentions (Go to Q2)
97 Refused (Go to Q2)
98 Don’t know (Go to Q2)
Base: If Q1 = 1
Q1a What denomination do you identify with?

DO NOT READ RESPONSES, CODE ALL THAT APPLY

[PROGRAMMER: ACCEPT MULTIPLE RESPONSES]

1 Adventist (Seventh Day Adventist)
2 Anglican/Episcopalian
3 Baptist (Southern Baptist Convention, Independent Baptist, National Baptist Convention)
4 Congregationalist (United Church of Christ)
5 Holiness (Church of the Nazarene, Free Methodist Church)
6 Lutheran (Evangelical Lutheran Church in America, Lutheran Church Missouri Synod)
7 Methodist (United Methodist Church, African Methodist Episcopal)
8 Nondenominational / Ecumenical
9 Pentecostal (Assemblies of God, Church of God in Christ, Church of God)
10 Presbyterian
11 Restorationist (Church of Christ, Disciples of Christ)
12 Something else (specify)
13 No preference
14 None in particular
96 No other mentions
97 (VOL) Refused
98 (VOL) Don’t know

Base: All
Q2 Aside from weddings and funerals, how often do you attend religious services?

1 More than once a week
2 Once a week
3 Once or twice a month
4 A few times a year
5 Seldom
6 Never
7 (VOL) Refused
8 (VOL) Don’t know

Base: All
Q3 What is your marital status?

1 Now married
2 Widowed
3 Divorced
4 Separated
5 Never married
7 (VOL) Refused
8 (VOL) Don’t know
**Base: All**

Q4  How many children do you have, including those no longer living with you?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 20

-2  (VOL) Don’t know
-1  (VOL) Refused

**Current Household Listing**

**Base: All**

Q5a  How many people currently live in your household, including yourself? Please also include any children, under the age of 18.

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 15

-2  (VOL) Don’t know
-1  (VOL) Refused

IF Q5a > 1 GO TO STATEMENT BEFORE Q5b THEN LOOP THROUGH Q5b AND Q5c WHERE THERE ARE Q5a-1 LOOPS

IF Q5a = -1, -2, OR 1 GO TO STATEMENT BEFORE Q6

Read text where Q5a > 1
I am going to ask you questions about every person who lives in your household—except you—starting with the oldest person.

**Base: Q5a > 1**

Q5b  How is the [1ST LOOP: oldest / OTHER LOOPS: next oldest] person in your household related to you?

DO NOT READ RESPONSES

1  Mother/Stepmother (Go to Q5c)
2  Father/Stepfather (Go to Q5c)
3  Spouse (Go to Q5c)
4  Sibling (Go to Q5c)
5  Birth Child (Go to Q5c)
6  Step Child (Go to Q5c)
7  Adopted Child (Go to Q5c)
8  Foster Child (Go to Q5c)
9  Grandchild (Go to Q5c)
10  Child (not specified) (Go to Q5c)
11  Niece/nephew (Go to Q5c)
12  Aunt/uncle (Go to Q5c)
13  Cousin (Go to Q5c)
14  Grandparent (Go to Q5c)
15  DELETED
16  In-law (Go to Q5c)
17  Other relative (Go to Q5c)
Q5c  How old is your [FILL FROM Q5b]?

NUMERIC ENTRY. RANGE = -2 TO 99.

-2  (VOL) Don’t know
-1  (VOL) Refused

If CATI_Q11 = 6 or 9 go to Q9
If CATI_Q11 ≠ 6 or 9 go to text below

Education and Training

Show text if CATI_Q11 ≠ 6 or 9
Now I’d like to talk about your educational background.

If CATI_Q11 = 1, 2, 5, or 7 go to Q7
If CATI_Q11 = 3 or 4 go to Q6
If CATI_Q11 = 8 go to Q8b

Base: If CATI_Q11 = 3 or 4
Q6  Do you have a high school diploma or a GED?

DO NOT READ RESPONSES

1  GED (Go to Q7)
2  High School Diploma (Go to Q9 if CATI_Q11=3; Go to Q7 if CATI_Q11=4)
3  Both (Go to Q7)
4  Neither (Go to Q7)
7  Refused (Go to Q7)
8  Don’t Know (Go to Q7)
**Base:** If \( CATI_{Q11} = 1, 2, 4, 5, 7 \) or \( Q6 = 1, 3, 4, 7, 8 \)

**Q7** What is the highest grade or year of regular school that you have completed and gotten credit for?

DO NOT READ RESPONSES

[DISPLAY ROWS AS FOLLOWS:]

IF \( CATI_{Q11} = 1 \): DISPLAY 1 TO 10, 97 TO 98  
IF \( CATI_{Q11} = 2 \): DISPLAY 11 TO 14, 97 TO 98  
IF \( Q6 = 1,3,4,7,8 \): DISPLAY 1 TO 18, 97 TO 98  
IF \( CATI_{Q11} = 4 \): DISPLAY 15 TO 18, 97 TO 98  
IF \( CATI_{Q11} = 5 \): DISPLAY 15 TO 18, 97 TO 98

1  No formal education (Go to Q9)  
2  Kindergarten (Go to Q9)  
3  Grade 1 (Go to Q9)  
4  Grade 2 (Go to Q9)  
5  Grade 3 (Go to Q9)  
6  Grade 4 (Go to Q9)  
7  Grade 5 (Go to Q9)  
8  Grade 6 (Go to Q9)  
9  Grade 7 (Go to Q9)  
10  Grade 8 (Go to Q9)  
11  Grade 9 (Go to Q9)  
12  Grade 10 (Go to Q9)  
13  Grade 11 (Go to Q9)  
14  Grade 12 (Go to Q9)  
15  1st year of college (If \( CATI_{Q11} = 5 \) go to Q8a; else go to Q9)  
16  2nd year of college (If \( CATI_{Q11} = 5 \) go to Q8a; else go to Q9)  
17  3rd year of college (If \( CATI_{Q11} = 5 \) go to Q8a; else go to Q9)  
18  4th year of college (If \( CATI_{Q11} = 5 \) go to Q8a; else go to Q9)  
19  1st year of graduate/professional school (Go to Q9)  
20  2nd year of graduate/professional school (Go to Q9)  
21  3rd year of graduate/professional school (Go to Q9)  
22  4th year of graduate/professional school (Go to Q9)  
23  5th year of graduate/professional school (Go to Q9)  
24  6th year of graduate/professional school (Go to Q9)  
25  7th year of graduate/professional school (Go to Q9)  
26  8th year of graduate/professional school (Go to Q9)  
27  9th year of graduate/professional school (Go to Q9)  
28  10th year of graduate/professional school (Go to Q9)  
97  Refused (If \( CATI_{Q11} = 5 \) Go to Q8a; else go to Q9)  
98  Don’t know (If \( CATI_{Q11} = 5 \) Go to Q8a; else go to Q9)
Base: If CATI_Q11 = 5
Q8a Is your Associate degree academic or is it for an occupation or vocation?

DO NOT READ RESPONSES

1  Academic (e.g., English, Math) (Go to Q9)
2  Occupation or vocation (e.g., veterinary assistant, dental hygienist) (Go to Q9)
7  Refused (Go to Q9)
8  Don’t know (Go to Q9)

Base: If CATI_Q11 = 8
Q8b What is the highest degree you have received?

DO NOT READ RESPONSES

1  Master’s degree (e.g., MA, MS, MBA)
2  Professional school degree (e.g., MD, JD, DDS, DVM)
3  Doctorate degree (e.g., Ph.D., Ed.D., DBA)
7  Refused
8  Don’t know

Base: All
Q9 Are you currently participating in any regular schooling or in some type of training program that lasts at least two weeks, and that is designed to help you find a job, improve your job skills, or learn a new job?

DO NOT READ RESPONSES

1  Yes (Go to Q9a)
2  No (Go to Q10)
7  Refused
8  Don’t know

Base: If Q9 = 1
Q9a What kind of schooling or training is that?

DO NOT READ RESPONSES

1  Regular schooling
2  General Equivalency Diploma (GED)
3  English as a Second Language
4  Computer training
5  Work study program
6  Certification or training in a health care field
7  Job search
8  Hospitality program
9  Auto repair
10  Childcare or education
11  Driving
12  Cosmetology
13  Remedial life skills
14 Accounting/Financial
15 Law/paralegal
16 Social Work
17 Construction/maintenance
18 Business/Management/Entrepreneurial
19 On the job training
20 Basic Job training
21 Certification in criminal justice
22 Other
97 Refused
98 Don’t know

Base: All
Q10 Do you qualify for any educational benefits such as tuition assistance through the Post-9/11 GI Bill?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
Q11 Have you ever served on active duty in the U.S. Armed Forces, Reserves or National Guard?

1 Never served in the military
2 Only on active duty for training in the Reserves or National Guard
3 Now on active duty
4 On active duty in the past, but not now
7 (VOL) Refused
8 (VOL) Don’t know
Employment

Now I’d like to ask you a few questions about any jobs you may have.

Base: All

Q13 Are you now employed full-time, part-time or not employed?

1 EMPLOYED FULL-TIME/PART-TIME (CONTINUE TO Q14)
2 NOT EMPLOYED (SKIP TO Q22)
97 (VOL) REFUSED (SKIP TO Q23)
98 (VOL) DON’T KNOW (SKIP TO Q23)

Base: If Q13 = 1

Q14 Last week, did you have more than one job, including part-time and weekend work?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: If Q13 = 1

Q15 How many hours per week do you usually work at your main job? By main job, we mean the one at which you usually work the most hours.

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -3 TO 168.

-2 (VOL) Don’t know
-1 (VOL) Refused
-3 (VOL) Hours vary from week to week

Base: If Q13 = 1

Q16 What kind of business or industry is this? (IF NEEDED: What do they make or do where you work?)

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused
Base: If \( Q13 = 1 \)

Q17  Is this business or organization mainly manufacturing, retail trade, selling things to other businesses, or something else?

DO NOT READ RESPONSES

1  Manufacturing
2  Retail Trade
3  Wholesale Trade
4  Something Else (Specify: [TEXT ENTRY])
7  Refused
8  Don’t know

Base: If \( Q13 = 1 \)

Q18  What kind of work do you do, that is, what is your occupation? For example, plumber, teacher, farmer.

TEXT ENTRY

-2  (VOL) Don’t know
-1  (VOL) Refused

Base: If \( Q13 = 1 \)

Q19  Including overtime pay, tips, and commissions, what are your usual earnings on your primary job, before taxes or other deductions?

PROGRAMMER: NUMERIC ENTRY. ALLOW 2 DECIMAL POINTS. RANGE -2 TO 999,999.00

-2  (VOL) Don’t know (Go to Q21)
-1  (VOL) Refused (Go to Q21)

Base: If \( Q13 = 1 \)

Q20  Is this hourly, weekly, bi-weekly, monthly or annually?

DO NOT READ RESPONSES

1  Hourly
2  Weekly
3  Bi-weekly
4  Monthly
5  Annually
6  Other (Specify: [TEXT ENTRY])
Base: If Q13 = 1

Q21 Now I’d like to ask a few questions about benefits that may be available at your job. Through your employer are you eligible for any of the following benefits? By eligible we mean the benefit is available for you now, even if you have decided to not receive it or have not needed it.

[PROGRAMMER: Loop through the following items]

1  Health insurance
2  Sick leave
3  Paid vacation

[PROGRAMMER: Response items for each item]

DO NOT READ RESPONSES

1  Yes (Go to Q37)
2  No (Go to Q37)
7  Refused (Go to Q37)
8  Don’t know (Go to Q37)

Base: If Q13 = 2

Q22 What is the main reason that you did not work for pay last week?

DO NOT READ RESPONSES

1  Retired
2  Disabled
3  Unable to Work
4  Has Job But Temporarily Absent
5  Couldn’t Find Any Work
6  Child Care Problems
7  Family Responsibilities
8  In School or Other Training
9  Waiting For a New Job to Begin
10  Unemployed/laid off
11  Pregnant
12  Caring for Sick
13  No education/skills
14  Volunteer Work
15  No job
16  No work permit
17  Seasonal employment
18  Transportation problem
19  Fired
20  Moving houses
21  Don’t want work
22  Quit
23  Has baby
24  Other
97  Refused
98  Don’t know
Base: If \( Q_{13} > 1 \)

Q23  Do you currently want a job, either full-time or part-time?

DO NOT READ RESPONSES

1  Yes/Maybe/It Depends (Go to Q23a)
2  No (Go to Q24)
3  Retired (Go to Q24)
4  Disabled (Go to Q24)
5  Unable to Work (Go to Q24)
6  Refused (Go to Q24)
7  Don’t know (Go to Q24)

Base: If \( Q_{23} = 1 \)

Q23a  Have you been doing anything to find work during the past four weeks?

DO NOT READ RESPONSES

1  Yes (Go to Q23b)
2  No (Go to Q24)
3  Refused (Go to Q24)
4  Don’t know (Go to Q24)

Base: If \( Q_{23a} = 1 \)

Q23b  What are all the things you have done to find work during the past four weeks?

[PROGRAMMER: MULTIPLE SELECT]

DO NOT READ RESPONSES. ENTER ALL THAT APPLY.

1  Contacted Employer(s)
2  Contacted Public Employment Agency Programs/Courses
3  Contacted Private Employment Agency
4  Contacted Friends or Relatives
5  Interviewed for a Job
6  Contacted School/University Employer Center
7  Sent Out Resumes/Filled out Applications
8  Checked Union/Professional Registers
9  Placed or Answered Ads
10  Looked at Ads Directly
11  Attended Job Training
12  Nothing
13  Other
96  No other mentions
97  Refused
98  Don’t know
**Base: If Q13 > 1**

Q24  Last week, could you have started a job if one had been offered?

**DO NOT READ RESPONSES**

1  Yes (Includes “maybe” under conditions on hours, wages or type of work) (Go to Q37)
2  No (Go to Q24a)
7  Refused (Go to Q37)
8  Don’t know (Go to Q37)

**Base: If Q24=2**

Q24a  Why is that?

[PROGRAMMER: MULTIPLE SELECT]

**DO NOT READ RESPONSES. ENTER ALL THAT APPLY**

1  Waiting for a new job to begin
2  Own temporary illness
3  Going to school
4  Childcare responsibilities
5  Family responsibilities
6  Transport problems
7  Moving houses
8  Mental/Physical Illness
9  Other
96  No other mentions
97  Refused
98  Don’t know

**Base: All**

Q37  What kind of health insurance or health care coverage do you have for yourself?

[INTERVIEWER: RESPONSE SHOULD BE THINGS LIKE “THROUGH MY EMPLOYER” OR “MEDICAID” WE DON’T NEED NAME OF INSURANCE COMPANY]

**TEXT ENTRY**

-2  (VOL) Don’t know
-1  (VOL) Refused

---

**Adult Income and Benefits**

**Base: All**
BEN1  Next I’d like to talk with you about any income or public assistance you (IF (Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10) and Q5c < 18 for any HH member rostered: or your child under the age of 18) may receive. Do you receive any form of public assistance or benefits from the government?

DO NOT READ RESPONSES

1  Yes (Go to Q36)
2  No (Go to INC1)
7  Refused (Go to Q36)
8  Don’t know (Go to Q36)

Q36  Are you (IF (Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10) and Q5c < 18 for any HH member rostered: or your child under the age of 18) now receiving help from the Supplemental Security Income program, called SSI?

DO NOT READ RESPONSES

1  Yes (IF Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10 go to Q36a;)
2  No (Go to Q38)
7  Refused (Go to Q38)
8  Don’t know (Go to Q38)

Base: If (Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10) and Q36 = 1
Q36a  Is the SSI for you or for your child?

If Q5b ≠ 5 and Q5b ≠ 6 and Q5bb ≠ 7 and Q5b ≠ 8 and Q5b ≠ 10 autopunch Q36a=1
If not [{Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10} and Q5c < 18] autopunch Q36a=1

[INTERVIEWER- IF NECESSARY: READ ANSWER CHOICES ALOUD]

1  You
2  Your child
3  Both you and your child
7  REFUSED
8  DON’T KNOW

Base: IF BEN1 = 1,7,8
Q38  Now I’d like to ask you about cash assistance some families receive on a regular basis. For example, they may get a monthly check. Some people call this assistance “welfare,” AFDC, TANF (“tan-eff”) or “public aid.” I’ll use the word “welfare.” Are you regularly receiving welfare benefits now?

DO NOT READ RESPONSES

1  Yes
2  No
7  Refused
8  Don’t know
**Base: IF BEN1 = 1,7,8**

Q39 Do you (IF Q5b = 5 or Q5b = 6 or Q5b = 7 or Q5b = 8 or Q5b = 10: or your child) receive food stamps or WIC (“wick”)?

DO NOT READ RESPONSES

1. Yes
2. No
7. Refused
8. Don’t know

**Base: If BEN1 = 2**

INC1 What was your total annual household income last year, in 2016, before taxes?

NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = 0 TO 999999.00.

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: If BEN1 = 1,7,8**

INC_Alt1 What was your total annual household income last year, in 2016, before taxes, as you reported to the government?

NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = 0 TO 999999.00.

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: IF ((INC1 = -1 or -2) or (INC_Alt1= -1 or -2))**

INC1a Could you please tell me if it was… (READ LIST)
1. Under $20,000 (ASK INC1b)
2. $20,000 to under $50,000, or (ASK INC1c)
3. $50,000 or more (ASK INC1d)
-1 (VOL) Don’t know
-2 (VOL) Refused

**Base: If INC1a=1**

INC1b Is it…(READ LIST)
1. Under $10,000, or
2. $10,000 to under $20,000?
-1 (VOL) Don’t know
-2 (VOL) Refused

**Base: If INC1a=2**

INC1c Is it…(READ LIST)
1. $20,000 to under $30,000
2. $30,000 to under $40,000, or
3. $40,000 to under $50,000?
**Base: If INC1a=3**

INC1d Is it…(READ LIST)

1. $50,000 to under $75,000
2. $75,000 to under $100,000
3. $100,000 to under $150,000, or
4. $150,000 or more?

-1 (VOL) Don’t know
-2 (VOL) Refused

**Base: If BEN1=2**

INC2 If you are not the only income earner in your household, what was your individual annual income last year, in 2016, before taxes?

NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = -3 TO 999999.00.

-3 (VOL) Only income earner in household
-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: If BEN1=1,7,8**

INC_Alt2 If you are not the only income earner in your household, what was your individual annual income last year, in 2016, before taxes, as you reported to the government?

NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = -3 TO 999999.00.

-3 (VOL) Only income earner in household
-2 (VOL) Don’t know
-1 (VOL) Refused
Base: IF ((INC2 = -1 or -2) or (INC_Alt2= -1 or -2))

INC2a Could you please tell me if it was… (READ LIST)
1. Under $20,000 (ASK INC1b)
2. $20,000 to under $50,000, or (ASK INC1c)
3. $50,000 or more (ASK INC1d)
-1  (VOL) Don’t know
-2  (VOL) Refused

Base: If INC2a=1

INC2b Is it…(READ LIST)
1. Under $10,000, or
2. $10,000 to under $20,000?
-1  (VOL) Don’t know
-2  (VOL) Refused

Base: If INC2a=2

INC2c Is it…(READ LIST)
1. $20,000 to under $30,000
2. $30,000 to under $40,000, or
3. $40,000 to under $50,000?
-1  (VOL) Don’t know
-2  (VOL) Refused

Base: If INC2a=3

INC2d Is it…(READ LIST)
1. $50,000 to under $75,000
2. $75,000 to under $100,000
3. $100,000 to under $150,000, or
4. $150,000 or more?
-1  (VOL) Don’t know
-2  (VOL) Refused

Base: All

Q25 Now I want to ask about accounts at a bank, savings and loan, or credit union. Please think about any accounts which you own, including joint accounts. Do you currently have any checking accounts, savings accounts, or any other type of bank account at a bank, credit union, or other financial institution? (IF NEEDED: Do not include retirement savings like 401(k), 403(b), and IRAs.)

DO NOT READ RESPONSES

1  Yes (Go to Q25a)
2  No (Go to Q25b)
7  Refused (Go to INC3)
8  Don’t know (Go to INC3)
**Base: If Q25 = 1**

Q25a  About how much is in these accounts all together?

PROGRAMMER: NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = -2 TO 999999.00. (Go to INC1)

-2   (VOL) Don’t know (Go to INC3)
-1   (VOL) Refused (Go to INC3)

**Base: If Q25 = 2**

Q25b  What is the most important reason you don’t have a bank account?

DO NOT READ RESPONSES

1   Don’t write enough checks to make it worthwhile
2   Do not have enough money to keep in an account
3   Don’t like dealing with banks
4   Service charges are too high
5   No bank has convenient hours or locations
6   Institutions will take their money due to debt
7   Bank closed account due to overdraft fees
8   Don’t have enough money
9   Other response
97  Refused
98  Don’t know

**Base: All**

INC3  In your own words, what would you say are the main factors that have contributed to your current financial situation? (PROBE: Up to 3 mentions; accept multiple responses)

Response 1 [TEXT ENTRY]
Response 2 [TEXT ENTRY]
Response 3 [TEXT ENTRY]

INC3DK   (VOL) Don’t know
INC3RF   (VOL) Refused

**Base: All**

Q26  Do you or does anyone in your household own a car or truck, or other motor vehicle that runs and can be driven on the road?

DO NOT READ RESPONSES

1   Yes (Go to Q26a)
2   No (Go to Q27)
7   Refused (Go to Q27)
8   Don’t know (Go to Q27)

**Base: If Q26 = 1**

Q26a  Thinking about the vehicles that you own, did you borrow money or get financing to purchase any of your vehicles?
**Base: All**

Q27  Do you or does anyone in your household have any unpaid medical bills?

**DO NOT READ RESPONSES**

1  Yes (Go to Q27a)
2  No (Go to Q28)
7  Refused (Go to Q28)
8  Don’t know (Go to Q28)

**Base: If Q27 = 1**

Q27a  About how much do you and your household still owe on your medical bills?

NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE = -2 TO 999999.00

-2  (VOL) Don’t know
-1  (VOL) Refused

**Base: All**

Q28  Do you have any credit or charge cards, including major credit cards like Visa or MasterCard, or charge cards from a store or gas station such as Sears or Mobil?

**DO NOT READ RESPONSES**

1  Yes (Go to Q28a)
2  No (Go to Q29)
7  Refused (Go to Q29)
8  Don’t know (Go to Q29)

**Base: If Q28 = 1**

Q28a  About how much do you and your household currently owe on all your credit and charge cards?

PROGRAMMER: NUMERIC ENTRY. ALLOW UP TO 2 DECIMAL POINTS. RANGE -2 TO 999999.00.

-2  (VOL) Don’t know
-1  (VOL) Refused

**Base: All**

Q29  Are debt collectors contacting you to ask that you make payments on any debts? (If needed: Not only your credit cards.)

**DO NOT READ RESPONSES**
Base: All
Q30 Do you have an overdue balance you’re paying down on your utilities?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
Q31 Are there any bills that you are ignoring or not making any payment on?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
Q32 If you needed to borrow $500 for three months, is there some person or place you could borrow it from?

DO NOT READ RESPONSES

1 Yes (Go to Q32a)
2 No (Go to Q33)
3 Would not borrow (Go to Q33)
7 Refused (Go to Q33)
8 Don’t know (Go to Q33)
Base: If Q32 = 1
Q32a Where would you go first to borrow $500 for three months?

DO NOT READ RESPONSES

1 Friends or family
2 A finance company
3 A payday loan at a check cashing outlet
4 Someone in my neighborhood who lends out money and charges interest
5 A community loan fund (or church loan fund)
6 A cash advance on my credit card
7 A bank (or savings bank, savings & loan, or credit union)
8 A pawn shop
9 A furniture store
10 Job/pension/union
11 I would not borrow
12 Other
97 Refused
98 Don’t know

Base: All
Q33 How often do you or your household put off buying something you need because you don't have money? Would you say all the time, frequently, occasionally, rarely, or never?

DO NOT READ RESPONSES

1 All the time
2 Frequently
3 Occasionally
4 Rarely
5 Never
7 Refused
8 Don’t know

If CATI_Q1A > 1 go to Q34
If CATI_Q1A = 1 go to Q35

Base: CATI_Q1A > 1
Q34 Did you ever live with your mother for a period of 6 months or more, at age 25 or older?

NOTE TO INTERVIEWER: MOTHER MAY BE BIOLOGICAL OR STEPMOTHER OR ADOPTIVE MOTHER

DO NOT READ RESPONSES

1 Yes (Go to Q34a)
2 No (Go to Q35)
6 No mother (Go to LT1)
97 Refused (Go to Q35)
98 Don’t know (Go to Q35)
Base: If $Q34 = 1$

Q34a Would you say that living with her was mainly to help your mother out, to help you out, or because it would be helpful to both of you?

DO NOT READ RESPONSES

1. To help your mother out
2. To help you out
3. Helpful to both
4. Neither
5. Refused
6. Don’t know

Base: All

Q35 Did any of your brothers or sisters ever live with your mother for a period of 6 months or more, at age 25 or older?

NOTE TO INTERVIEWER: SIBLINGS MAY BE BIOLOGICAL OR STEP-SIBLINGS OR ADOPTED SIBLINGS OR HALF SIBLINGS

DO NOT READ RESPONSES

1. Yes (Go to Q35a)
2. No (Go to LT1)
6. No brothers or sisters / Brothers or sisters all younger than age 25 (Go to LT1)
7. Refused (Go to LT1)
8. Don’t know (Go to LT1)

Base: If $Q35 = 1$

Q35a Would you say that your brother or sister living with your mother was mainly to help your mother out, to help your brother or sister out, or because it would be helpful to both of them?

DO NOT READ RESPONSES

1. To help your mother out
2. To help your brother or sister out
3. Helpful to both
4. Neither
97. Refused
98. Don’t know
Lifetime Traumas

Base: All
LT1  Now I’d like to ask you a few questions about events you may have experienced. For each of the following events, please indicate whether the event occurred at any point in your life...

[PROGRAMMER: Loop through following items]

1  Have you ever been homeless?
2  Have you ever been in a major fire, flood, earthquake, or other natural disaster?
3  Did you ever have a life-threatening illness or accident?
4  Did your spouse or a child of yours ever have a life-threatening illness or accident?
5  Have you ever fired a weapon in combat or been fired upon in combat?
6  Has your spouse, partner, or child ever been addicted to drugs or alcohol?
7  Were you the victim of a serious physical attack or assault in your life?
8  Has a child of yours ever died?

[PROGRAMMER: Response options for each item.]

DO NOT READ RESPONSES

1  Yes
2  No
7  Refused
8  Don’t know

Base: All
LT2  Have you ever been arrested by the police or taken into custody for an illegal or delinquent offense? Do not include arrests for minor traffic violations.

DO NOT READ RESPONSES

1  Yes (Go to LT3)
2  No (Go to LT4)
7  Refused (Go to LT4)
8  Don’t know (Go to LT4)

Base: If LT2 = 1
LT3  In total, how many times have you been arrested?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2, -1, 1 TO 999.

-2  (VOL) Don’t know
-1  (VOL) Refused
Lifetime Traumas Before the Age of 18

Base: All
LT4 Before you were 18 years old, did you ever have to do a year of school over again?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
LT5 Before you were 18 years old, were you ever in trouble with the police?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
LT6 Before you were 18 years old, did either of your parents drink or use drugs so often that it caused problems in the family?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
LT7 Before you were 18 years old, were you ever physically abused by either of your parents?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know
Adult Health

Show text to all
Now I’d like to ask you some questions about your general health.

Base: All
HEAL1 In general, how is your health: excellent, very good, good, fair, or poor?

DO NOT READ RESPONSES

1  Excellent
2  Very good
3  Good
4  Fair
5  Poor
7  Refused
8  Don’t know

Base: All
HEAL2 In general, how has your health been in the past month: excellent, very good, good, fair, or poor?

DO NOT READ RESPONSES

1  Excellent
2  Very good
3  Good
4  Fair
5  Poor
7  Refused
8  Don’t know

Base: All
HEAL3 In the past year, have you had a routine physical examination?

DO NOT READ RESPONSES

1  Yes
2  No
7  Refused
8  Don’t know
Base: All
HEAL4 How tall are you?

If respondent only knows height in meters:

<table>
<thead>
<tr>
<th>Feet</th>
<th>Inches</th>
<th>Meters</th>
<th>Centimeters</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>1.52</td>
<td>152</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.55</td>
<td>155</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1.57</td>
<td>157</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1.60</td>
<td>160</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>1.63</td>
<td>163</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1.65</td>
<td>165</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>1.68</td>
<td>168</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>1.70</td>
<td>170</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>1.73</td>
<td>173</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>1.75</td>
<td>175</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1.78</td>
<td>178</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>1.80</td>
<td>180</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1.83</td>
<td>183</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1.85</td>
<td>185</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1.88</td>
<td>188</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1.91</td>
<td>191</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1.93</td>
<td>193</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1.96</td>
<td>196</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1.98</td>
<td>198</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>2.01</td>
<td>201</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>2.03</td>
<td>203</td>
</tr>
</tbody>
</table>

Feet: NUMERIC ENTRY. INTEGERS ONLY. RANGE = 3 TO 8.
Inches: NUMERIC ENTRY. INTEGERS ONLY. RANGE = 0 TO 11.

HEAL4DK (VOL) Don’t know
HEAL4RF (VOL) Refused
Base: All
HEAL5 How much do you weigh? (IF NEEDED: Your best guess is fine.)

If respondent only knows weight in kilograms:

120 pounds = 54 kilograms
130 pounds = 59 kilograms
140 pounds = 64 kilograms
150 pounds = 68 kilograms
160 pounds = 73 kilograms
170 pounds = 77 kilograms
180 pounds = 82 kilograms
190 pounds = 86 kilograms
200 pounds = 91 kilograms
210 pounds = 95 kilograms
220 pounds = 100 kilograms
230 pounds = 104 kilograms
240 pounds = 109 kilograms
250 pounds = 113 kilograms

Pounds: NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO -1, 50 TO 600.

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All
HEAL6 Have you ever suffered from depression, anxiety, emotional distress, or mental illness of any kind?

DO NOT READ RESPONSES

1 Yes (Go to HEAL7)
2 No (Go to DIS1)
7 Refused (Go to DIS1)
8 Don’t know (Go to DIS1)
**Base: If HEAL6 = 1**

HEAL7 What was the illness?

DO NOT READ RESPONSES. ALLOW MULTIPLE RESPONSES.

[PROGRAMMER: MULTIPLE SELECT]

1 Depression
2 Anxiety
3 Bipolar
4 Schizophrenia
5 Paranoia
9 ADHD
10 PTSD
11 Eating Disorder
12 Emotional distress
13 Other
96 No other mentions
97 Refused
98 Don’t know

**Discrimination**

Now I have a few questions about discrimination. Sometimes people feel like they are discriminated against, or treated badly or differently because of their race or ethnicity.

**Base: All**

DIS1 Can you think of one or more occasions in the last 6 months when you felt you were treated unfairly because of your race or ethnicity in the following places?

[PROGRAMMER: Loop through following items]

1 At your school or work?
2 In a store where you were shopping or a restaurant where you wanted to eat?
3 When you met someone for the first time?
4 In dealing with the police?

[PROGRAMMER: Response options for each item.]

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

*Show text for all*

Sometimes people feel they are discriminated against, or treated badly or differently because they might not have quite as much money as other people, or because of the way they dress or talk.

---

1 From MTO.
Can you think of one or more occasions in the last 6 months when you felt you were treated unfairly because of how much money your family has or the way you dress or talk?

[PROGRAMMER: Loop through following items]

1. At your school or work?
2. In a store where you were shopping or a restaurant where you wanted to eat?
3. When you met someone for the first time?
4. In dealing with the police?

[PROGRAMMER: Response options for each item.]

DO NOT READ RESPONSES

1. Yes
2. No
7. Refused
8. Don’t know

Discount Rates

Show text to all
Suppose that after having helped a friend with some small jobs, they offer to send you a small amount of money in return for your help. They tell you that they can either send you something now, or send you a little more if you are willing to wait one month. If they pay you now, they will put $40 in the mail tomorrow. If they pay you one month from now, they will send you slightly more than that. Suppose that you trust them to pay you what they promise, when they promise it, and that either payment is equally convenient for them.

Base: All
DSC1  Would you rather they mailed you $40 tomorrow or $47 one month from now?

DO NOT READ RESPONSES

1. $40 tomorrow (Go to DSC4)
2. $47 one month from now (Go to DSC2)
7. Refused (Go to DSC2)
8. Don’t know (Go to DSC2)

Base: If DSC1 ≠ 1
DSC2  Now suppose the choice were between $40 now and $45 one month from now. Would you rather they mailed you $40 tomorrow or $45 one month from now?

DO NOT READ RESPONSES

1. $40 tomorrow (Go to statement before RISK1)
2. $45 one month from now (Go to DSC3)
7. Refused (Go to DSC3)
8. Don’t know (Go to DSC3)
Base: If DSC2 ≠ 1
DSC3  Now suppose the choice were between $40 now and $42 one month from now. Would you rather they mailed you $40 tomorrow or $42 one month from now?

DO NOT READ RESPONSES

1  $40 tomorrow (Go to statement before RISK1)
2  $42 one month from now (Go to statement before RISK1)
7  Refused (Go to statement before RISK1)
8  Don’t know (Go to statement before RISK1)

Base: If DSC1 = 1
DSC4  Now suppose the choice were between $40 now and $50 one month from now. Would you rather they mailed you $40 tomorrow or $50 one month from now?

DO NOT READ RESPONSES

1  $40 tomorrow (Go to DSC5)
2  $50 one month from now (Go to statement before RISK1)
7  Refused (Go to statement before RISK1)
8  Don’t know (Go to statement before RISK1)

Base: If DSC4 = 1
DSC5  Now suppose the choice were between $40 now and $55 one month from now. Would you rather they mailed you $40 tomorrow or $55 one month from now?

DO NOT READ RESPONSES

1  $40 tomorrow
2  $55 one month from now
7  Refused
8  Don’t know

Risk Aversion

Show text to all
Suppose you have a choice between two, equally good jobs that lasted for one month. The first would pay you $600 for the month. The second job would pay you an amount that depends on how the company as a whole did during that month. It is possibly better paying, but your earnings will be less certain.
Base: All
RISK1 There is a 50-50 chance that the second job will pay $1200, and a 50-50 chance it will pay $400. Which would you choose -- the job that pays $600 for sure, or the job with an equal chance of paying either $1200 or $400?

DO NOT READ RESPONSES

1 $600 for sure (Go to RISK4)
2 Equal chance of paying either $1200 or $400 (Go to RISK2)
7 Refused (Go to RISK2)
8 Don’t know (Go to RISK2)

Base: If RISK1 ≠ 1
RISK2 Now suppose there is a 50-50 chance that the second job will pay $1200, and a 50-50 chance that it will pay $300. Which would you choose -- the job that pays $600 for sure, or the job with an equal chance of paying either $1200 or $300?

DO NOT READ RESPONSES

1 $600 for sure (Go to RES)
2 Equal chance of paying either $1200 or $300 (Go to RISK3)
7 Refused (Go to RISK3)
8 Don’t know (Go to RISK3)

Base: If RISK2 ≠ 1
RISK3 Now suppose there is a 50-50 chance that the second job will pay $1200, and a 50-50 chance that it will pay $150. Which would you choose -- the job that pays $600 for sure, or a job with an equal chance of paying either $1200 or $150?

DO NOT READ RESPONSES

1 $600 for sure (Go to RES)
2 Equal chance of paying either $1200 or $150 (Go to RES)
7 Refused (Go to RES)
8 Don’t know (Go to RES)

Base: If RISK1 = 1
RISK4 Now suppose there is a 50-50 chance that the second job will pay $1200, and a 50-50 chance that it will pay $480. Which would you choose -- the job that pays $600 for sure, or a job with an equal chance of paying either $1200 or $480?

DO NOT READ RESPONSES

1 $600 for sure (Go to RISK5)
2 Equal chance of paying either $1200 or $480 (Go to RES)
7 Refused (Go to RES)
8 Don’t know (Go to RES)
If $RISK4 = 1$

Now suppose there is a 50-50 chance that the second job will pay $1200, and a 50-50 chance that it will pay $540. Which would you choose -- the job that pays $600 for sure, or a job with an equal chance of paying either $1200 or $540?

DO NOT READ RESPONSES

1 $600 for sure
2 Equal chance of paying either $1200 or $540
7 Refused
8 Don’t know

Brief Resilience Scale²

Base: All

I am going to read you a series of sentences. For each of these, please tell me how much you agree with each sentence. There are no right or wrong answers.

[PROGRAMMER: Loop through the following items.]

1 I tend to bounce back quickly after hard times
2 I have a hard time making it through stressful events
3 It does not take me long to recover from a stressful event
4 It is hard for me to snap back when something bad happens
5 I usually come through difficult times with little trouble
6 I tend to take a long time to get over set-backs in my life

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #1

1 Strongly disagree
2 Disagree
3 Neutral
4 Agree
5 Strongly agree
7 (VOL) Refused
8 (VOL) Don’t know

² Smith et. al 2008
Childhood Experience

*Base: All*

CHHH1 Who did you live with for the majority of the time at age 9?

DO NOT READ RESPONSES, CODE ALL THAT APPLY. INTERVIEWER IF NEEDED: FOSTER PARENT SHOULD BE NOTED IN OTHER SPECIFY.

(PROGRAMMER: ACCEPT MULTIPLE RESPONSES]

1. Mother/Stepmother
2. Father/Stepfather
3. Grandparent(s)
4. Younger brother(s)/stepbrother(s)
5. Younger sister(s)/stepsister(s)
6. Older brother(s)/stepbrother(s)
7. Older sister(s)/stepsister(s)
8. Twin/triplet/etc. brother(s)
9. Twin/triplet/etc. sister(s)
10. Niece/nephew(s)
11. Aunt/uncle(s)
12. Cousin(s)
13. Other (Specify: [TEXT BOX])
96. No other mentions
97. Refused
98. Don’t know

*Base: If CHHH1_4 selected.*

CHHH2 How many younger brothers or stepbrothers did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

*Base: If CHHH1_5 selected.*

CHHH3 How many younger sisters or stepsisters did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused
Base: If CHHH1_6 selected.
CHHH4 How many older brothers or stepbrothers did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: If CHHH1_7 selected.
CHHH5 How many older sisters or stepsisters did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: If CHHH1_8 selected.
CHHH6 How many brothers born at the same time as you did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: If CHHH1_9 selected.
CHHH7 How many sisters born at the same time as you did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: If CHHH1_10 selected.
CHHH8 How many nieces and nephews did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused
**Base: If CHHH1_11 selected.**

CHHH9 How many uncles and aunts did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: If CHHH1_12 selected.**

CHHH10 How many cousins did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: If CHHH1_13 selected.**

CHHH11 How many [FILL WITH TEXT FROM CHHH1_13txt] did you live with at age 9? (If needed: For a majority of the time.)

[NUMERIC ENTRY. RANGE = -2-9]

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: All**

NB1 How many times did you move between the ages of 5 and 17?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-12.
INTERVIEWER: ENTER 12 FOR “12 OR MORE”

0 (Go to NB4)
1-12 (Go to NB2)
-2 (VOL) Don’t know (Go to NB4)
-1 (VOL) Refused (Go to NB4)

[PROGRAMMER: Construct read-in variable NB1txt from NB1 where 1 = “1st”, 2 = “2nd”, 3 = “3rd”, 4 = “4th”, 5 = “5th. ONLY PROGRAM 5 LOOPS.]

[PROGRAMMER: If NB1 = 1 to 5, loop through mentions 1 to 5 in NB2 and NB3. For each loop NB2 then NB3 should be asked before moving to the next loop iteration.]

**Base: Loop iteration from NB1**

NB2 How old were you when you moved the [FILL WITH NB1txt] time?

NUMERIC ENTRY, ACCEPT MULTIPLE ANSWERS. INTEGERS ONLY. RANGE = -2, -1, 5-17.

-2 (VOL) Don’t know
-1 (VOL) Refused
Base: Loop iteration from NB1

NB3  What was the main reason you moved at age [INSERT AGE FROM NB2]?

TEXT ENTRY

98  (VOL) Don’t know
99  (VOL) Refused

Base: All

NB4  Please think about the neighborhood that you spent the most amount of time in when you were under the age of 18. I’d like to ask some questions about that neighborhood. For the following, the answer choices are “Very safe, Safe, Unsafe, Very unsafe.”

[PROGRAMMER: Loop through following items]

1  How safe were the parking lots and sidewalks near your neighborhood school?
2  How safe did you feel at home alone at night?
3  How safe were the streets near your home during the day?
4  How safe were the streets near your home at night?

[PROGRAMMER: Response options for each item]

USE SHOWCARD #2

1  Very safe
2  Safe
3  Unsafe
4  Very unsafe
7  (VOL) Refused
8  (VOL) Don’t know

Base: All

NB5  The next questions ask about problems in your neighborhood. Your answer choices are: Big problem, Small problem, No problem at all. In your neighborhood, how bad of a problem was...

[PROGRAMMER: Loop through following items]

1  Litter or trash on the streets or sidewalks?
2  Graffiti or writing on the walls?
3  People drinking in public?

[PROGRAMMER: Response options for each item]

USE SHOWCARD #3

1  Big problem
2  Small problem
3  No problem at all
7  (VOL) Refused
8  (VOL) Don’t know
Can you remember any of the following happening in your neighborhood?

[PROGRAMMER: Loop through following items]

1. Was anyone’s purse, wallet, or jewelry snatched from them?
2. Was anyone threatened with a knife or gun?
3. Was anyone beaten or assaulted?
4. Was anyone stabbed or shot?
5. Did anyone try to break into your home?

[PROGRAMMER: Response options for each item]

1. Yes
2. No
7. (VOL) Refused
8. (VOL) Don’t know

If a group of neighborhood children were skipping school and hanging out on a street corner, how likely is it that your neighbors would do something about it? Would you say very likely, likely, unsure, unlikely, or very unlikely?

DO NOT READ RESPONSES

1. Very likely
2. Likely
3. Unsure (includes don’t know)
4. Unlikely
5. Very unlikely
7. Refused

If some children were spray-painting graffiti on a local building, how likely is it that your neighbors would do something about it? Would you say very likely, likely, unsure, unlikely, or very unlikely?

DO NOT READ RESPONSES

1. Very likely
2. Likely
3. Unsure (includes don’t know)
4. Unlikely
5. Very unlikely
7. Refused
Finances

Read to all
Now think about your family when you were growing up, from birth until age 16.

Base: All
CFIN1 Would you say your family during that time was pretty well off financially, about average, or poor?

DO NOT READ RESPONSES

1 Pretty well off
2 About average
3 Poor
7 Refused
8 Don’t know

Base: All
CFIN2 While you were growing up, from birth until age 16, did financial difficulties ever cause you or your family to move to a different place?

DO NOT READ RESPONSES

1 Yes (Go to CFIN2a)
2 No (Go to CFIN3)
7 Refused (Go to CFIN3)
8 Don’t know (Go to CFIN3)

Base: If CFIN2 = 1
CFIN2a How old were you at the time?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 16.

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All
CFIN3 From birth until age 16, was there a time when you or your family received help from relatives because of financial difficulties?

DO NOT READ RESPONSES

1 Yes (Go to CFIN3a)
2 No (Go to CFIN4)
7 Refused (Go to CFIN4)
8 Don’t know (Go to CFIN4)

---

3 Mostly HRS, some MTO Baseline,
Base: If CFIN3 = 1
CFIN3a How old were you at the time?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 16.

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All
CFIN4 From birth until age 16, was there a time of several months or more when your father had no job?

DO NOT READ RESPONSES

1 Yes (Go to CFIN4a)
2 No (Go to CFIN5)
6 Did not have father / Father deceased (Go to CFIN6)
8 Refused (Go to CFIN5)
7 Don’t know (Go to CFIN5)

Base: If CFIN4 = 1
CFIN4a How old were you at the time?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 16.

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: CFIN4≠6
CFIN5 What was your father’s occupation when you were age 16?

TEXT ENTRY

-4 (VOL) Father was unemployed
-3 (VOL) Did not have father / Father deceased
-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All
CFIN6 Did your mother ever get AFDC (IF NEEDED: “Aid to Families with Dependent Children”) or welfare from your birth until age 16?

DO NOT READ RESPONSES

1 Yes (Go to CFIN6a)
2 No (Go to CFIN7)
6 Did not have mother / Mother deceased (Go to CFIN8)
7 Refused (Go to CFIN7)
8 Don’t know (Go to CFIN7)
Base: If CFIN6 = 1
CFIN6a How old were you at the time?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 16.

7   (VOL) Refused
8   (VOL) Don’t know

PROGRAMMER: SKIP CFIN7 IF CFIN4=6 OR CFIN6=6
Base: IF CFIN4≠6 AND CFIN6≠6
CFIN7 Did you live with both of your parents until you were 16?

DO NOT READ RESPONSES

1   Yes
2   No (includes one or both parents deceased / some of the time)
7   Refused
8   Don’t know

Base: All
CFIN8 Did you ever live in the same household with a grandparent for a year or more until you were 16?

DO NOT READ RESPONSES

1   Yes
2   No
7   Refused
8   Don’t know

PROGRAMMER: SKIP CFIN9 IF CFIN6=6
Base: CFIN6≠6
CFIN9 What portion of the time did your mother work outside the home until you were 16: all of the time, some of the time, or not at all?

DO NOT READ RESPONSES

1   All of the time
2   Some of the time
3   Not at all
4   Mother deceased or not around
7   Refused
8   Don’t know
Base: All
CFIN10 How often do you remember the following happening to your parents or guardian until you were age 16? Would you say always, often, sometimes, rarely, or never? [INSERT STATEMENT FROM LOOP SHOWN BELOW]

[PROGRAMMER: Loop through the following statements in CFIN10 and CFIN10a. CFIN10 is always asked. CFIN10a is only asked if CFIN10 ≤ 3. CFIN10a should be asked immediately after CFIN10 for each item.]

1. Being unable to find child care or being forced to take a child out of child care because they couldn’t pay?
2. Falling behind in rent or mortgage payments?
3. Falling behind in gas, electric, or phone bills?
4. Being unable to pay for adequate transportation to get to work or school?
5. Being unable to get medical care because of the cost?
6. Having trouble paying a credit card balance?
7. Having too little money to buy enough food?
8. Being a victim of a crime?
9. Having a problem with alcohol or drug abuse?

USE SHOWCARD #4

1. Always
2. Often
3. Sometimes
4. Rarely
5. Never
6. (VOL) Refused
7. (VOL) Don’t know

Base: If the loop iteration of CFIN10 ≤ 3
CFIN10a What ages were you at the time?

[“INTERVIEWER: ENTER 0 FOR AGE IF HAPPENED SINCE BIRTH”]

Start age: NUMERIC ENTRY. RANGE = -2-16 [CFIN10a1]
End age: NUMERIC ENTRY. RANGE = -2-16 [CFIN10a2]

7. (VOL) Refused
8. (VOL) Don’t know

PROGRAMMER: CHECK CFIN10a1 ≤ CFIN10a2. IF CFIN10a1 > CFIN10a2 DISPLAY “End age should be greater than or equal to start age.”
**Base: All**

CFIN11 From your birth until age 16, did your parents (or guardian) divorce or separate due to financial problems?

DO NOT READ RESPONSES

1  Yes (Go to CFIN11a)
2  No (Go to CFIN12)
7  Refused (Go to CFIN12)
8  Don’t know (Go to CFIN12)

**Base: If CFIN11 = 1**

CFIN11a How old were you at the time? (If necessary: How old were you the first/last time it happened

-2  (VOL) Don’t know
-1  (VOL) Refused

**Parents Incarceration**

**Base: All**

CFIN12 Was there ever a time during your childhood that a parent (or guardian) living with you had to serve time in jail or prison?

DO NOT READ RESPONSES

1  Yes (Go to CFIN12a)
2  No (Go to REL1)
7  Refused (Go to REL1)
8  Don’t know (Go to REL1)

**Base: If CFIN12 = 1**

CFIN12a How many years of your childhood was your parent (or guardian) in jail? [IF NEEDED: By childhood, we mean before the age of 18; if more than one sentence, please add the total time, if less than one year, code 0]

Gave response RANGE 0-18

-2  (VOL) Don’t know
-1  (VOL) Refused

**Base: If CFIN12 = 1**

CFIN12b How old were you when your parent (or guardian) went to jail? [IF NEEDED: if more than one sentence, we want the first one]

Start age:  NUMERIC ENTRY. RANGE = -2-18 [CFIN12b1]

-2  (VOL) Don’t know

---

4 From National Survey of Children’s Health 2011-12 adapted.
Quality of Early Life Relationships

PROGRAMMER: SKIP ITEM 1 IF CFIN6=6 / SKIP ITEM 2 IF CFIN4=6
Base: IF CFIN4#6 AND CFIN6#6

REL1 I am going to read you a series of sentences about your early relationships with your parents. For each of these, please tell me how much you agree with each sentence. There are no right or wrong answers.

[PROGRAMMER: Loop through the following items.]

1 I had a good relationship with my mother before age 18.
2 I had a good relationship with my father before age 18.

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #5

1 Strongly disagree
2 Disagree
3 Neither agree nor disagree
4 Agree
5 Strongly agree
6 (VOL) No parent / Parent deceased
7 (VOL) Refused
8 (VOL) Don’t know

---

5 HRS.
I am going to read you a series of sentences. For each of these, please tell me if you would say a lot, some, a little, or not at all. There are no right or wrong answers.

[PROGRAMMER: Loop through the following items.]

1. How much time and attention did your mother give you when you needed it?
2. How much effort did your mother put into watching over you and making sure you had a good upbringing?
3. How much did your mother teach you about life?

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #6

1. A lot
2. Some
3. A little
4. Not at all
6. (VOL) No mother / Mother deceased
7. (VOL) Refused
8. (VOL) Don’t know

When you were a child, were there any adults in your life that you felt you could depend on 100% of the time?

DO NOT READ OPTIONS

1. Yes (Go to REL4)
2. No (Go to PAR1)
7. Refused (Go to PAR1)
8. Don’t know (Go to PAR1)

How many of these people were there?

NUMERIC ENTRY. INTEGERS ONLY RANGE = -2, -1, 1 TO 99.

-2. (VOL) Don’t know
-1. (VOL) Refused
Base: If REL3 = 1
REL5 Can you please describe how each of these people was related or known to you?

[PROGRAMMER: MULTIPLE RESPONSE]

DO NOT READ RESPONSES, CODE ALL THAT APPLY

1 Mother/Stepmother
2 Father/Stepfather
3 Sibling
4 Aunt/uncle
5 Cousin
6 Grandparent
7 Teacher
8 Family friend
9 Religious leader
10 Coach
11 Other: specify (TEXT ENTRY)
12 Other unrelated person
96 No other mentions
97 Refused
98 Don’t Know
Parental Behavior\(^6\)

*Base: All*  
PAR\(1\) I am going to read you a series of sentences. For each of these, please tell me how often each of these happened to you. There are no right or wrong answers. Please think about when you were between the ages of 5 and 12.

[IF NEEDED: Whoever was most like a parent to you.]  
[PROGRAMMER: Loop through the following items.]

1. How often would your parent say that you did something that gave him/her pleasure and enjoyment?
2. How often would your parent say that you did something that greatly irritated him/her and got on his/her nerves?
3. How often did your parent read to you?
4. How often did your parent physically punish you as a child, for example by a spanking?
5. How often did your parent praise you as a child, by saying something like “Good for you!” “What a nice thing you did!” “Thank you!” or “That's good going!”
6. How often did your parent tell you about his/her experience, by saying something like, “I saw a pretty bird outside just a little while ago,” or “I exercised so hard that I got really tired,” or “I was able to give some directions today to somebody that got lost,” or “I really like the way the sky looks now.”
7. How often did you and your parent talk or play with each other, focusing attention on each other for five minutes or more, without your parent asking or telling you to do anything?
8. How often did your parent tell you to do something, with an irritated or angry tone of voice?
9. How often did you and your parent engage in make-believe play, where you each played the part of a character, and together made up a story to act out with each other?
10. How often did you laugh with your parent?
11. How often did your parent yell or speak to you in a very loud voice, with irritated or angry emotion?

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #7

1. Never
2. Less than once a week
3. About once a week
4. About three or four times a week
5. About once a day
6. Several times each day
7. Many times a day
96. (VOL) No parental figure
97. (VOL) Refused
98. (VOL) Don’t know

\(^6\) Rephrased from Parent Practices Survey.
Base: All
PAR2 Thinking about when you were between the ages of 5 and 12…

[IF NEEDED: Whoever was most like a parent to you.]

[PROGRAMMER: Loop through the following items.]

1 What fraction of days did you get three meals, one in the morning, one around noon, and one in the evening?
2 What fraction of days did you get a bath or shower at one particular time, known as your bath time?
3 What fraction of days did you eat all of the following: some meat (or other high protein food), some fruits or vegetables, some milk products, and some bread or grain products?
4 When you and your parent set out to do something fun together, what fraction of the time did it actually turn out to be fun?
5 What fraction of days was your parent too worn out and exhausted to do something fun with you?
6 How often do you think the thought went through your parent’s mind that he wished he didn’t have to spend so much time with you?
7 Think of all the times that your parent commented to you about your behavior. What fraction were congratulation or approval?
8 Think of all the times that your parent commented to you about your behavior. What fraction were correction or disapproval?

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #8

1 Never
2 Some, but less than a quarter of the time
3 Between a quarter and half the time
4 Not all the time, but more than three quarters of the time
5 All the time
6 (VOL) No parental figure
7 (VOL) Refused
8 (VOL) Don’t know
Base: All
PAR3 What fraction of the time did you go to bed at one particular time, known as your bedtime when you were age 9?

[IF NEEDED: Whoever was most like a parent to you.]

1. There was no regular or official bedtime
2. There was an official bedtime, but it was never kept
3. There was an official bedtime and it was kept some, but less than a quarter of the time
4. There was an official bedtime and it was kept between a quarter and half the time
5. There was an official bedtime and it was kept between half and three quarters of the time
6. There was an official bedtime and it was kept not all the time, but more than three quarters of the time
7. There was an official bedtime and it was kept all the time
96 (VOL) No parental figure
97 (VOL) Refused
98 (VOL) Don’t know

Base: All
PAR4 Suppose when you were 9, you were handling an object that your parent definitely did not want you to handle. Suppose he or she told you to put the object down, and you defiantly said “No!” Of the following options, which would your parent respond with most of the time?

[IF NEEDED: Whoever was most like a parent to you.]

1. Spank you
2. Send you to your room for half an hour or more
3. Yell at you
4. Repeat the request until you obeyed
5. Ignore you
6. Send you to a room for two to five minutes
7. Send you to a room for five to thirty minutes
8. Show some disapproval in his/her voice and in his/her face, and physically get the object from you, and from then on, if possible, keep the object in a place you couldn’t reach
9 (VOL) Something else (Specify: [TEXT ENTRY])
96 (VOL) No parental figure
97 (VOL) Refused
98 (VOL) Don’t know

Base: All
PAR5 Did your parent keep you from seeing television shows and movies that had a lot of violence or meanness in them between the ages of 5 and 12? [IF NEEDED: Whoever was most like a parent to you.]

DO NOT READ RESPONSES

1. Yes
2. No
6 (VOL) No parental figure
7 Refused
8. Don’t know
**Base: All**

**PAR6** How often did you see adults or teenagers in your house physically fighting with or hitting or otherwise trying to hurt each other when you were between the ages of 5 and 12? Would you say…

[IF NEEDED: Whoever was most like a parent to you.]

1 Never
2 Less than once a week
3 About once a week
4 About three or four times a week
5 About once a day
6 Several times each day
7 Many times each day
96 (VOL) No parental figure
97 (VOL) Refused
98 (VOL) Don’t know

**Base: All**

**PAR7** When your parent gave you a command or order to do something, what fraction of the time did he or she make sure that you did it?

[IF NEEDED: Whoever was most like a parent to you.]

1 Never
2 Some, but less than a quarter of the time
3 Between a quarter and half the time
4 Between half and three quarters of the time
5 Not all the time, but more than three quarters of the time
6 All the time
96 (VOL) No parental figure
97 (VOL) Refused
98 (VOL) Don’t know

**Base: All**

**PAR8** Did your parents arrange the objects in your house so that those things that they didn’t want you to mess with were not within your reach, so that they didn’t have to command you to stay out of them? Would you say…

[IF NEEDED: Whoever was most like a parent to you.]

1 Many things were in reach that a child should leave alone
2 A good number of things were in reach that a child should leave alone
3 A few things were in reach that a child should leave alone
4 Almost no things were in reach that a child should leave alone
5 No things were in reach that a child should leave alone
6 (VOL) No parental figure
7 (VOL) Refused
8 (VOL) Don’t know

49

299
Base: All
PAR9  Thinking about when you were ages 5 to 12…

[PROGRAMMER: Loop through following items]

1. How often were you able to get your way by having a tantrum?
2. How often did your parent tell you he or she may leave you if you didn’t behave better?
3. How often were you punished for crying?
4. How often were you punished for wetting yourself?
5. How often did your parent or someone else tell you that you are bad or not as good as someone else?
6. How often did you see an adult in the house raise his voice in anger at some other adult in the house?
7. How often did you see an adult in the house do something kind, friendly, or very much appreciated by another adult in the house?

[PROGRAMMER: Response options for each item]

USE SHOWCARD #9

1. Never
2. Less than once a week
3. About once a week
4. About three or four times a week
5. About once a day
6. Several times each day
7. Many times each day
96  (VOL) No parental figure
97  (VOL) Refused
98  (VOL) Don’t know

Base: All
PAR10  When you asked your parent a question, what fraction of the time did he or she answer it in an enthusiastic and interested way, rather than an irritated way? Would you say your parent…

1. Never answered enthusiastically
2. Answered enthusiastically some, but less than a quarter of the time
3. Between a quarter and half the time
4. Between half and three quarters of the time
5. Not all the time, but more than three quarters of the time
6. Answered enthusiastically all the time
96  (VOL) No parental figure
97  (VOL) Refused
98  (VOL) Don’t know
Childhood Health Questions

Show text to all
Consider your health while you were growing up, before you were 16 years old.

Base: All
CHE1 Would you say that your health during that time was excellent, very good, good, fair, or poor?

DO NOT READ RESPONSES

1 Excellent
2 Very good
3 Good
4 Fair
5 Poor
6 Refused
8 Don’t know

Base: All
CHE2 When you were growing up, before you were 16 years old, did you miss a month or more of school because of a health problem?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

Base: All
CHE3 Before you were 16 years old, did you have any of the following childhood diseases?

[PROGRAMMER: Loop through following items]

1 Measles
2 Mumps
3 Chicken Pox
4 Asthma
5 Diabetes
6 Epilepsy or seizures

[PROGRAMMER: Response items for each option]

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

7 HRS.
**Base: All**

**CHE4** Before you were 16 years old, did you have difficulty seeing, even with eye glasses or prescription lenses?

**DO NOT READ RESPONSES**

1. Yes
2. No
3. Refused
4. Don’t know

**Base: All**

**CHE5** Did your parents or guardians smoke during your childhood?

**DO NOT READ RESPONSES**

1. Yes (including some or some of the time)
2. No
3. Refused
4. Don’t know

**Base: All**

**CHE6** Before you were 16 years old, did you have...

[PROGRAMMER: Loop through following items]

1. Bronchitis, wheezing, hay fever, shortness of breath or sinus infection or something else that made it difficult to breathe
2. A speech impairment (IF NEEDED: A speech impairment means having difficulty speaking.)
3. Allergies
4. Heart trouble
5. Chronic ear problems or infections (IF NEEDED: Chronic means lasting for a long time)
6. Severe headaches or migraines
7. Stomach problems
8. High blood pressure
9. Depression
10. Drug or alcohol problems
11. Any other emotional or psychological problems

[PROGRAMMER: Response items for each option]

**DO NOT READ RESPONSES**

1. Yes
2. No
3. Refused
4. Don’t know
**Base: All**

**CHE7** Before you were 16 years old, did you have a blow to the head, a head injury or head trauma that was severe enough to require medical attention, or to leave you unconscious or to cause memory loss for a period of time?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know

**Base: All**

**CHE8** Before you were 16 years old, were you ever disabled for six months or more because of a health problem? That is, were you unable to do the usual activities of classmates or other children your age?

DO NOT READ RESPONSES

1 Yes (Go to CHE8a)
2 No (Go to CHE9)
7 Refused (Go to CHE9)
8 Don’t know (Go to CHE9)

**Base: If QCHE_8 = 1**

**CHE8a** What was the cause of that disability?

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused

**Base: All**

**CHE9** In grade school or high school, did you ever have a problem in learning the usual lessons, and had to regularly attend special classes, receive special training sessions, or attend a different school for more than six months?

DO NOT READ RESPONSES

1 Yes
2 No
7 Refused
8 Don’t know
Base: All
CHE10 Were there any other important or serious health problems that you had before age 16 that you could tell me about?

TEXT ENTRY

-3 (VOL) No such problem
-2 (VOL) Don’t know
-1 (VOL) Refused

Childhood Risky Attitudes and Behaviors

Base: All
RATT1 Please think back to when you were 16. I am going to read a list of items that describe feelings or thoughts people sometimes have. Please tell me if at the age of 16, you would have said that the item is true or often true, sometimes true, or not true of you.

[PROGRAMMER: Loop through following items]

1 I had trouble concentrating or paying attention
2 I lied or cheated
3 I teased others a lot
4 I disobeyed my parents
5 I had trouble sitting still
6 I had a hot temper
7 I would rather have been alone than with others
8 I hung around with kids who got into trouble
9 I disobeyed at school
10 I didn’t get along with other kids
11 I had trouble getting along with teachers

[PROGRAMMER: Response items for each option]

SHOWCARD #10

1 True
2 Often true
3 Sometimes true
4 Not true
7 (VOL) Refused
8 (VOL) Don’t know

Show text to all
Please answer the next few questions based on your behavior as a teenager.
Base: All
RATT2 First, I would like to ask you about smoking habits. As a teenager, did you smoke cigarettes?

DO NOT READ RESPONSES

1  Yes (Go to RATT3)
2  No (Go to RATT5)
7  Refused (Go to RATT5)
8  Don’t know (Go to RATT5)

Base: If RATT2 = 1
RATT3 At the age of 16, how many days in a month did you smoke a cigarette?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 31

-2  (VOL) Don’t know
-1  (VOL) Refused

Base: If RATT2 = 1
RATT4 When you smoked a cigarette at that time, how many cigarettes did you usually smoke each day? (Note to interviewers: Assume a pack contains 20 cigarettes.)

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 99

-2  (VOL) Don’t know
-1  (VOL) Refused

Show text to all
Next I would like to ask you some questions about drinking alcoholic beverages, including beer, wine, or liquor.

Base: All
RATT5 As a teenager, did you ever have a drink of an alcoholic beverage? By a drink we mean a can or bottle of beer, a glass of wine, a mixed drink, or a shot of liquor. Do not include childhood sips that you might have had from an older person’s drink.

DO NOT READ RESPONSES

1  Yes (Go to RATT6)
2  No (Go to RATT8)
7  Refused (Go to RATT8)
8  Don’t know (Go to RATT8)

Base: If RATT5 = 1
RATT6 At the age of 16, on how many days in a month did you have one or more drinks of an alcoholic beverage?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 31

-2  (VOL) Don’t know
-1  (VOL) Refused
Base: If RATT5 = 1
RATT7 On the days that you drank alcohol, about how many drinks did you usually have?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 99
-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All
RATT8 As a teenager, did you ever use marijuana (that is grass or pot)?

DO NOT READ RESPONSES
1 Yes (Go to RATT9)
2 No (Go to RATT10)
7 Refused (Go to RATT10)
8 Don’t know (Go to RATT10)

Base: If RATT8 = 1
RATT9 At the age of 16, on how many days in a month did you use marijuana?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 31
7 (VOL) Refused
8 (VOL) Don’t know

Base: All
RATT10 Excluding marijuana and alcohol, as a teenager, did you ever use any other drugs like cocaine or crack or heroin, or any other substance not prescribed for you by a doctor, in order to get high or to achieve an altered state?

DO NOT READ RESPONSES
1 Yes (Go to RATT11)
2 No (Go to text before PFRI1)
7 Refused (Go to text before PFRI1)
8 Don’t know (Go to text before PFRI1)

Base: If RATT10 = 1
RATT11 About how many times did you use any of these drugs or other substances in a year?

NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2 TO 365
-2 (VOL) Don’t know
-1 (VOL) Refused
Social Networks of Parents

Show text to all
Now I’d like to ask you a few questions about any friendships your parents or guardians had while you were growing up.

Base: All
PFRI1 About how many close friends did your parents have when you were a child? These are people they felt at ease with, could talk to about private matters, or call on for help. Would you say that your parents had no close friends, one or two, three to five, six to ten, or more than ten?

DO NOT READ RESPONSES

1. No close friends (Go to SUC1)
2. 1 or 2 (Go to PFRI2)
3. 3 to 5 (Go to PFRI2)
4. 6 to 10 (Go to PFRI2)
5. More than 10 (Go to PFRI2)
6. Refused (Go to SUC1)
7. Don’t know (Go to SUC1)

Base: If PFRI1 = 2, 3, 4, 5
PFRI2 About how many of your parents’ close friends…

[PROGRAMMER: LOOP THROUGH FOLLOWING ITEMS]

1. Lived in the same neighborhood as you
2. Graduated from college
3. Worked full-time
4. Were a different race or ethnicity than you

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #11

1. All
2. Most
3. Some
4. A few
5. None
6. (VOL) Refused
7. (VOL) Don’t know

---

Adapted from MTO.
Ideas of Success

SUC1 In a few words, what is your idea of success?

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All

SUC2 What was your idea of success when you were young?

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All

SUC3 When you were young, did you believe you would grow up to be successful?

DO NOT READ RESPONSES

1 Yes (Go to SUC4)
2 No (Go to SUC4)
7 Refused (Go to SUC7)
8 Don’t know (Go to SUC7)

Base: If SUC3 = 1,2

SUC4 Did you ever (IF SUC3=1: stop / IF SUC3=2: start) believing you would be successful?

DO NOT READ RESPONSES

1 Yes (Go to SUC5)
2 No (Go to SUC7)
7 Refused (Go to SUC7)
8 Don’t know (Go to SUC7)
Base: If $SUC_4 = 1$

$SUC_5$ At what age did you (IF $SUC_3 = 1$: stop / IF $SUC_3 = 2$: start) believing you would be successful?

NUMERIC ENTRY. RANGE = -2-99

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: If $SUC_4 = 1$

$SUC_6$ Why did you (IF $SUC_3 = 1$: stop / IF $SUC_3 = 2$: start) believing you would be successful?

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: All

$SUC_7$ Was there a time in your life when help could have made all the difference?

DO NOT READ RESPONSES

1 Yes
2 No (Go to SELF)
7 Refused (Go to SELF)
8 Don’t know (Go to SELF)

Base: $SUC_7 = 1$

$SUC_8$ At what age was that time in your life that help could have made all the difference?

NUMERIC ENTRY. RANGE = -2-99

-2 (VOL) Don’t know
-1 (VOL) Refused

Base: $SUC_7 = 1$

$SUC_9$ What type of help would have made the difference?

TEXT ENTRY

-2 (VOL) Don’t know
-1 (VOL) Refused
Brief Self-Control Scale

Base: All

SELF  I am going to read you a series of statements. For each of these, please indicate how much the statement reflects how you typically are on a 1 to 5 scale where 1 means “Not at all” and 5 means “Very much”.

[PROGRAMMER: Loop through the following items.]

1 I am good at resisting temptation.
2 I have a hard time breaking bad habits.
3 I am lazy.
4 I say inappropriate things
5 I do certain things that are bad for me, if they are fun.
6 I refuse things that are bad for me.
7 I wish I had more self-discipline.
8 People would say that I have iron self-discipline.
9 Pleasure and fun sometimes keep me from getting work done.
10 I have trouble concentrating.
11 I am able to work effectively toward long-term goals.
12 Sometimes I can’t stop myself from doing something, even if I know it is wrong.
13 I often act without thinking through all the alternatives.

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #12

1 1 - Not at all
2 2
3 3
4 4
5 5 - Very much
7 (VOL) Refused
8 (VOL) Don’t know

Rotter’s Locus of Control Scale

Show text to all

I am going to read you a series of pairs of statements. For each pair, please select the statement that is closer to your opinion. In addition, tell me whether the statement you select is much closer to your opinion or slightly closer. In some cases, you may find that you believe both statements; in other cases you may believe neither one. Even when you feel this way about a pair of statements, select the one statement which is more nearly true in your opinion. Try to consider each pair of statements separately when making your choices; do not be influenced by your previous choices.
Base: All
LOC1 “What happens to me is my own doing” or “Sometimes I feel that I don’t have enough control over the direction my life is taking.”

[RANDOMIZE ORDER OF STATEMENTS IN STEM]

1. What happens to me is my own doing.
2. Sometimes I feel that I don’t have enough control over the direction my life is taking.
7. (VOL) Refused
8. (VOL) Don’t know

LOC1a Would you say this statement is much closer to your opinion or slightly closer?
1. Much closer
2. Slightly closer

Base: All
LOC2 “When I make plans, I am almost certain that I can make them work” or “It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.”

[RANDOMIZE ORDER OF STATEMENTS IN STEM]

1. When I make plans, I am almost certain that I can make them work.
2. It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune anyhow.
7. (VOL) Refused
8. (VOL) Don’t know

LOC2a Would you say this statement is much closer to your opinion or slightly closer?
1. Much closer
2. Slightly closer

Base: All
LOC3 “In my case getting what I want has little or nothing to do with luck” or “Many times we might just as well decide what to do by flipping a coin.”

[RANDOMIZE ORDER OF STATEMENTS IN STEM]

1. In my case getting what I want has little or nothing to do with luck.
2. Many times we might just as well decide what to do by flipping a coin.
7. (VOL) Refused
8. (VOL) Don’t know

LOC3a Would you say this statement is much closer to your opinion or slightly closer?
1. Much closer
2. Slightly closer

Base: All
LOC4 “Many times I feel that I have little influence over the things that happen to me” or “It is impossible for me to believe that chance or luck plays an important role in my life.”

[RANDOMIZE ORDER OF STATEMENTS IN STEM]
Many times I feel that I have little influence over the things that happen to me.
It is impossible for me to believe that chance or luck plays an important role in my life.
(VOL) Refused
(VOL) Don’t know

LOC4a Would you say this statement is much closer to your opinion or slightly closer?
1 Much closer
2 Slightly close
Dweck Mindset Instrument

Base: All
MIND  I am going to read you a series of sentences. For each of these, please indicate how much you agree with the sentence.

[PROGRAMMER: Loop through the following items.]

1  You have a certain amount of intelligence, and you really can’t do much to change it.
2  Your intelligence is something about you that you can’t change very much.
6  You can learn new things, but you can’t really change your basic intelligence.

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #13

1  Strongly Agree
2  Agree
3  Mostly Agree
4  Mostly Disagree
5  Disagree
6  Strongly Disagree
7  (VOL) Refused
8  (VOL) Don’t know

CAGE-AID

Read following statement to all
I am going to read you a series of questions. For each of these, please answer Yes or No. When thinking about drug use, include illegal drug use and the use of prescription drug other than prescribed.

[PROGRAMMER: Loop through the following items. AID1 will be asked first. If AID1 = 1, AID2 will be asked. AID2 will always follow AID1 for each iteration of the loop where AID1 = 1.]

1  Have you ever felt that you ought to cut down on your drinking or drug use?
2  Have people annoyed you by criticizing your drinking or drug use?
3  Have you ever felt bad or guilty about your drinking or drug use?
4  Have you ever had a drink or used drugs first thing in the morning to steady your nerves or get rid of a hangover?

Base: All
AID1  [Statement from loop]

DO NOT READ OPTIONS

1  Yes (Ask AID2)
2  No (Go to next loop iteration)
7  Refused
8  Don’t know
Base: AID1 = 1
AID2 At what age?
   IF MULTIPLE EPISODES, RECORD FIRST AND MOST RECENT EPISODES.

Earliest  NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [AID2_1]
Most recent NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [AID2_2]

[PROGRAMMER: Check that AID2_2 is either missing or AID2_2 ≥ AID2_1.]

-2  (VOL) Don’t know
-1  (VOL) Refused

GAD 7

Base: All
GAD I am going to read you a series of problems. For each of these, please indicate how often you have been bothered by it over the last 2 weeks.

[PROGRAMMER: Loop through the following items.]

1  Feeling nervous, anxious or on edge
2  Not being able to stop or control worrying
3  Trouble relaxing
4  Being so restless that it is hard to sit still
5  Becoming easily annoyed or irritable
6  Feeling afraid as if something awful might happen

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #14

1  Not at all
2  Several days
3  More than half the days
4  Nearly every day
7  (VOL) Refused
8  (VOL) Don’t know

Base: All
ANX Have you ever in your life had an attack of fear or panic when all of a sudden you felt very frightened, anxious, or uneasy?

DO NOT READ RESPONSES

1  Yes
2  No
7  Refused
8  Don’t know
**Patient Health Questionnaire**

*Base: All*

PHQ1  I am going to read you a series of problems. For each of these, please indicate how often you have been bothered by it over the past two weeks.

[PROGRAMMER: Loop through the following items.]

1. Little interest or pleasure in doing things
2. Feeling down, depressed or hopeless
3. Trouble falling asleep, staying asleep or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself – or that you’re a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed. Or, the opposite – being so fidgety or restless that you have been moving around a lot more than usual
9. Thoughts that you would be better off dead or of hurting yourself in some way.

[PROGRAMMER: Response options for each item]

USE SHOWCARD #15

1. Not at all
2. Several days
3. More than half the days
4. Nearly every day
7. (VOL) Refused
8. (VOL) Don’t know

*Base: If any of PHQ1_1 to PHQ1_9 between 2 and 4*

PHQ2  How difficult have those problems made it for you to do your work, take care of things at home, or get along with other people?

1. Not difficult at all
2. Somewhat difficult
3. Very difficult
4. Extremely difficult
7. (VOL) Refused
8. (VOL) Don’t know
Base: All

PHQ3 Have you ever in your life had an episode lasting several days or longer when most of the day you felt sad, empty or depressed?

DO NOT READ RESPONSES

1  Yes (Go to PHQ4)
2  No (Go to PHQ5)
7  Refused (Go to PHQ5)
8  Don’t know (Go to PHQ5)

Base: If PHQ3 = 1

PHQ4 At what age?

IF MULTIPLE EPISODES, RECORD FIRST AND MOST RECENT EPISODES.

Earliest NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [PHQ4_1]
Most recent NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [PHQ4_2]
-2  (VOL) Don’t know
-1  (VOL) Refused

[PROGRAMMER: Check that PHQ4_2 is either missing or PHQ4_2 ≥ PHQ4_1.]

Base: All

PHQ5 Have you ever had an episode lasting several days or longer when you lost interest in most things you usually enjoy like work, hobbies, and personal relationships?

DO NOT READ RESPONSES

1  Yes (Go to PHQ6)
2  No (Go to FAM1)
7  Refused (Go to FAM1)
8  Don’t know (Go to FAM1)

Base: If PHQ5 = 1

PHQ6 At what age?

IF MULTIPLE EPISODES, RECORD FIRST AND MOST RECENT EPISODES.

Earliest NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [PHQ6_1]
Most recent NUMERIC ENTRY. INTEGERS ONLY. RANGE = -2-99. [PHQ6_2]
-2  (VOL) Don’t know
-1  (VOL) Refused

[PROGRAMMER: Check that PHQ6_2 is either missing or PHQ6_2 ≥ PHQ6_1.]
Family Environment Scale

**Base: All**

FAM1 I am going to read you some statements about families. For each of these, please indicate whether they were true of the family you lived with while you were between the ages of 5 and 12. You may feel that some of the statements were true for some family members and false for others. Say True if the statement was true for most members. Say False if the statement was false for most members. If the members are evenly divided, decide what was the stronger overall impression and answer accordingly. Remember, we would like to know what your family seemed like to you. So do not try to figure out how other members would have seen your family.

[PROGRAMMER: Loop through the following items.]

1. We fought a lot in our family
2. Family members attended church, [IF Q1_4=1 OR Q1_8=1 OR Q1_9=1: temple / IF Q1=7: mosque / ELSE: synagogue], or another place of worship or Sunday School fairly often
3. Family members were rarely ordered around
4. Family members rarely became openly angry
5. We didn’t say prayers in our family
6. There were very few rules to follow in our family
7. Family members sometimes got so angry they threw things
8. We often talked about the religious meaning of Christmas, [IF Q1=7: Ramadan / IF Q1=8: Vesak / IF Q1=9: Diwali / ELSE: Passover], or other holidays
9. There was one family member who makes most of the decisions
10. Family members hardly ever lost their tempers
11. We didn’t believe in heaven or hell
12. There were set ways of doing things at home
13. Family members often criticized each other
14. Family members had strict ideas about what is right or wrong
15. There was a strong emphasis on following rules in our family
16. Family members sometimes hit each other
17. We believed that there were some things you just have to take on faith
18. Everyone had an equal say in family decisions
19. If there was a disagreement in our family, we tried hard to smooth things over and keep the peace
20. In our family each person had different ideas about what was right and wrong
21. We could do whatever we wanted to in our family
22. Family members often tried to one-up or out-do each other
23. The Bible [IF Q1_4=1 OR Q1_7=1 OR Q1_8=1 OR Q1_9=1: or another holy book like] [IF Q1_4=1: the Book of Mormon] [IF Q1_7=1: the Koran] [IF Q1_8=1: the Sutras] [IF Q1_9=1: the Vedas] was a very important book in our home
24. Rules were pretty inflexible in our household
25. In our family, we believed you don’t ever get anywhere by raising your voice
26. Family members believed that if you sinned, you would be punished
27. You couldn’t get away with much in our family
[PROGRAMMER: Response options for items]

1   True
2   False
7   (VOL) Refused
8   (VOL) Don’t know

Base: All
FAM2 When you answered these questions, who did you have in mind?

[PROGRAMMER: MULTIPLE RESPONSE]

CODE ALL THAT APPLY

1   Your brothers and sisters
2   Your parent
3   Your roommate
4   Other family members
5   Your spouse or partner
6   Your child or children, or
7   Someone else?
96  (VOL) No other mentions
97  (VOL) Refused
98  (VOL) Don’t know

50-item IPIP version of the Big Five Markers

Base: All
PIP I am going to read you a series of statements about yourself. For each of these, please indicate whether it is 1. Very Inaccurate, 2. Moderately Inaccurate, 3. Neither Accurate Nor Inaccurate, 4. Moderately Accurate, or 5. Very Accurate as a description of you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age.

[Interviewer: If respondent has trouble with the answer choices, please assist them]

[PROGRAMMER: Loop through the following items.]

1   I am the life of the party.
2   I feel little concern for others.
3   I am always prepared.
4   I get stressed out easily.
5   I have a rich vocabulary.
6   I don’t talk a lot.
7   I am interested in people.
8   I leave my belongings around.
9   I am relaxed most of the time.
10  I have difficulty understanding abstract ideas.
11  I feel comfortable around people.
12  I insult people.
13  I pay attention to details.
14  I worry about things.
15  I have a vivid imagination.
16  I keep in the background.
17  I sympathize with others’ feelings.
18  I make a mess of things.
19  I seldom feel blue.
20  I am not interested in abstract ideas.
21  I start conversations.
22  I am not interested in other people’s problems.
23  I get chores done right away.
24  I am easily disturbed.
25  I have excellent ideas.
26  I have little to say.
27  I have a soft heart.
28  I often forget to put things back in their proper place.
29  I get upset easily.
30  I do not have a good imagination.
31  I talk to a lot of different people at parties.
32  I am not really interested in others.
33  I like order.
34  I change my mood a lot.
35  I am quick to understand things.
36  I don’t like to draw attention to myself.
37  I take time out for others.
38  I ignore my duties.
39  I have frequent mood swings.
40  I use difficult words.
41  I don’t mind being the center of attention.
42  I feel others’ emotions.
43  I follow a schedule.
44  I get irritated easily.
45  I spend time reflecting on things.
46  I am quiet around strangers.
47  I make people feel at ease.
48  I am exacting in my work.
49  I often feel blue.
50  I am full of ideas.

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #16

1  Very Inaccurate
2  Moderately Inaccurate
3  Neither Accurate Nor Inaccurate
4  Moderately Accurate
5  Very Accurate
6  (VOL) Refused
7  (VOL) Don’t know
Short Grit Scale

Base: All
GRIT I am going to read you a number of statements that may or may not apply to you. For the most accurate score, when responding, think of how you compare to most people — not just the people you know well, but most people in the world. There are no right or wrong answers, so just answer honestly!

[PROGRAMMER: Loop through the following items.]

1. New ideas and projects sometimes distract me from previous ones.
2. Setbacks don’t discourage me.
3. I have been obsessed with a certain idea or project for a short time but later lost interest.
4. I am a hard worker.
5. I often set a goal but later choose to pursue a different one.
6. I have difficulty maintaining my focus on projects that take more than a few months to complete.
7. I finish whatever I begin.
8. I am diligent

[PROGRAMMER: Response options for each item.]

USE SHOWCARD #17

1. Very much like me
2. Mostly like me
3. Somewhat like me
4. Not much like me
5. Not like me at all
7. (VOL) Refused
8. (VOL) Don’t know

Rosenberg Self-Esteem Scale

Base: All
RSE I am going to read you a series of statements dealing with your general feelings about yourself. For each of these, please indicate how strongly you agree or disagree.

[PROGRAMMER: Loop through the following items.]

1. On the whole, I am satisfied with myself.
2. At times, I think I am no good at all.
3. I feel that I have a number of good qualities.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I certainly feel useless at times.
7. I feel that I’m a person of worth, at least on an equal plane with others.
8. I wish I could have more respect for myself.
9. All in all, I am inclined to feel that I am a failure.
10. I take a positive attitude toward myself.
Adverse Childhood Experience Questionnaire

*Base: All*

We are almost at the end of the survey. The questions I am about to read could be distressing to individuals with a history of post-traumatic stress disorder, or survivors of abuse and violence. You may find these questions upsetting. Please feel free to ask to skip these questions or discontinue the interview without penalty. You will not need to explain why you do not wish to answer or stop the interview. Remember that I also gave you a list of local support organizations at the beginning of the interview. I am going to read you a series of questions about your childhood. For each of these, please answer yes or no.

[PROGRAMMER: Loop through the following items.]

1. Did a parent or other adult in the household *often* swear at you, insult you, put you down, or humiliate you *or* act in a way that made you afraid that you might be physically hurt?
2. Did a parent or adult in the household *often* push, grab, slap, or throw something at you *or ever* hit you so hard that you had marks or were injured?
3. Did an adult or person at least 5 years older than you *ever* touch or fondle you or have you touch their body in a sexual way *or* try to actually have oral, anal, or vaginal sex with you when you were a minor? (IF NEEDED: A minor is someone under the age of 18.)
4. Did you *often* feel that no one in your family loved you or thought you were important or special *or* your family didn’t look out for each other, feel close to each other, or support each other?
5. Did you *often* feel that you didn’t have enough to eat, had to wear dirty clothes, and had no one to protect you *or* your parents were too drunk or high to take care of you or take you to the doctor if you needed it?
6. Were your parents *ever* separated or divorced?
7. Was your mother or stepmother *often* pushed, grabbed, slapped, or had something thrown at her *or sometimes or often* kicked, bitten, hit with a fist, or hit with something hard *or ever* repeatedly hit over at least a few minutes or threatened with a gun or knife?
8. Did you live with anyone who was a problem drinker or alcoholic or who used street drugs?
9. Was a household member depressed or mentally ill or did a household member attempt suicide?
10. Did a household member go to prison?
DO NOT READ RESPONSES

1  Yes
2  No
7  Refused
8  Don’t know
Closing

Base: All
PAYINFO Those are all the questions I have. Thanks so much. In order to document that you received your $150 VISA gift card for participating, I just need to record your full name:

FULL NAME

INTERVIEWER: RECORD GIFT CARD NUMBER HERE:

Thank you. My Supervisor or another staff member may call you to check that I talked to you today. It will only take a few minutes and would be a big help if you speak with them.

Have a great day!
### B.7. Literature Review of the Correlates of Intergenerational Mobility

<table>
<thead>
<tr>
<th>Authors</th>
<th>Data</th>
<th>Sample</th>
<th>Correlates</th>
</tr>
</thead>
<tbody>
<tr>
<td>McKernan and Ratcliffe (2005)</td>
<td>Monthly, longitudinal data from the 1988, 1990, and 1996 panels of the Survey of Income and Program Participation (SIPP) and monthly unemployment rates from the U.S. Department of Labor (2001) with quarterly real gross domestic product from the U.S. Department of Commerce (2001).</td>
<td>Sample size for each panel ranges from 14,000 to 36,700 households and data is collected for the preceding four months in each interview of the SIPP participant.</td>
<td>Education</td>
</tr>
<tr>
<td>Haskins (2008)</td>
<td>Panel Study of Income Dynamics (PSID) that tracks the mobility of adult children by comparing their income at roughly age 40 with that of their parents at about the same age.</td>
<td>All participants of the PSID survey where parents’ income was averaged over the period 1967-1971 and adult childrens’ incomes were averaged over selected years between 1995 and 2002.</td>
<td>Education</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
<td>Sample Details</td>
<td>Analysis Focus</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Metzler et al. (2017)</td>
<td>2010 Behavioral Risk Factor Surveillance System (BRFSS) that collects data on Adverse Childhood Experiences, education attainment, unemployment and poverty status.</td>
<td>Data from 27,834 noninstitutionalized adults surveyed from 10 states and the District of Columbia that used the adverse childhood experiences (ACE) module in the 2010 Behavioral Risk Factor Surveillance System.</td>
<td>Adverse Childhood Experiences</td>
</tr>
<tr>
<td><strong>Duncan and Rodgers (1988)</strong></td>
<td>1968-1982 waves of the Panel Study of Income Dynamics that contains data on family background, educational attainment and income.</td>
<td>1075 children who were under the age of 4 in 1968 who had not left home before the end of a 15-year period and for whom a full 15 years of family data were available.</td>
<td>Family events</td>
</tr>
<tr>
<td><strong>Sharkey and Torrats-Espinosa (2015)</strong></td>
<td>Equality of Opportunity project that contains data on intergenerational economic mobility, crime and community zone demographics.</td>
<td>287 urban commuting zones for which authors have non-missing data on crime and economic mobility.</td>
<td>Violent crime</td>
</tr>
<tr>
<td><strong>Freeman (1985)</strong></td>
<td>1979-80 National Bureau of Economic Research-Mathematica survey of Inner City youth and the 1979-81 National Longitudinal Survey of Young Men with data on church going habits, time allocations, school, work activity, frequency of socially deviant activity, family background and poverty status.</td>
<td></td>
<td>Church-going</td>
</tr>
</tbody>
</table>
Keels et al. (2005)

Information on the Gautreaux program participants were provided by the Leadership Council. This information contained mother’s age, AFDC recipiency status and number of children. Addresses were taken from a credit reporting service and the Illinois Department of Human Services Integrated Client Database Records. Neighborhood characteristic were calculated from the 1980 and 1990 U.S. censuses and crime data for neighborhoods were taken from FBI’s UCR records.

1506 randomly chosen families who moved as part of the Gautreaux program prior to 1990.

Neighborhood
<table>
<thead>
<tr>
<th>Source</th>
<th>Data Description</th>
<th>Sample Size</th>
<th>Data Relocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanbonmatsu et al. (2012)</td>
<td>Interviews of adults from “Moving to Opportunity” (MTO) households 10 to 15 years after families were randomized into the program. Data includes adult’s health and economic circumstances, physical measurements and blood samples.</td>
<td>3273 adults from MTO households.</td>
<td>Neighborhood</td>
</tr>
<tr>
<td>Chetty, Hendren and Katz (2016)</td>
<td>“Moving to Opportunity” (MTO) Data linked to federal income tax returns</td>
<td>4604 families in 5 large US cities that were randomized as part of the MTO experiment.</td>
<td>Neighborhood</td>
</tr>
</tbody>
</table>
## B.8. Literature Review of the Causal Impacts of Important Variables

<table>
<thead>
<tr>
<th>Correlate</th>
<th>Authors</th>
<th>Data</th>
<th>Sample</th>
<th>Causal Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Childhood</td>
<td>Belfield et al. (2006),</td>
<td>Numerous measures ranging from economic, criminal, and educational outcomes as well as cognition and personality between the ages of 3 and 40.</td>
<td>123 HighScope Perry Preschool Program participants who mostly consist of low-IQ, disadvantaged African-American children living in Ypsilanti, Michigan.</td>
<td>Monthly income at age 27 increases by $867 (Heckman, Pinto and Savelyev, 2013).</td>
</tr>
<tr>
<td>Education</td>
<td>Heckman et al. (2010),</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Heckman, Pinto and Savelyev (2013)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early Childhood</td>
<td>Barnett and Masse (2007),</td>
<td>Family background characteristics at study entry, school assessment test scores, college enrollment, crime, personality, health, behavior and earnings.</td>
<td>104 study participants and their families who were sampled to be economically disadvantaged were randomly assigned to the Abecedarian program.</td>
<td>The program effect on the gross earnings of future generations was estimated at $5700 (Barnett and Masse, 2007). Yearly labor incomes (in 2014 USD) increased by $3,578 for females and $17,214 for males. (Elango et al., 2015).</td>
</tr>
<tr>
<td>Education</td>
<td>Elango et al. (2015)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Early Childhood Education</td>
<td>Garces, Thomas and Currie (2002) and Elango et al. (2015)</td>
<td>Panel Study of Income Dynamics which records participation in Head Start and have long term follow up data. Data includes cognitive traits, high school completion, college attendance, crime, health behaviors and earnings.</td>
<td>Almost 4000 adults aged 18 and older in 1995 who were PSID respondents.</td>
<td>Earnings between the ages of 23 and 40 increased by 5.1 percentage points.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Education</td>
<td>Card (1999)</td>
<td>Survey of literature on the causal relationship between education and earnings. Selected studies use institutional aspects of the education system to form instrumental variables; studies of earnings and schooling of twins; and studies that explicitly model sources of heterogeneity in the returns to education.</td>
<td>An additional year of schooling can increase wages by 2-11%.</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------</td>
<td>-----------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Big 5 Personality Traits</td>
<td>Nyhus and Pons (2005)</td>
<td>CentER Saving Survey (CSS) which provides individual labor market details as well as responses to the Five Factor Personality Inventory (FFPI)</td>
<td>888 workers aged 16-65 who were part of the respondent pool for CentER Saving Survey in the Netherlands</td>
<td>Agreeableness has a significant and negative effect in the male sample while emotional stability positively affects the wage setting for women.</td>
</tr>
<tr>
<td>Big 5 Personality Traits</td>
<td>Heckman, Pinto and Savelyev (2013)</td>
<td>Numerous measures ranging from economic, criminal, and educational outcomes as well as cognition and personality between the ages of 3 and 40.</td>
<td>123 HighScope Perry Preschool Program participants who mostly consist of low-IQ, disadvantaged African-American children living in Ypsilanti, Michigan.</td>
<td>The Perry Pre School treatment increased monthly income at age 27 by $867. 20 percent of this treatment effect is explained by early improvements of personality traits.</td>
</tr>
<tr>
<td>Child Abuse</td>
<td>Currie and Widom (2010)</td>
<td>Using a prospective cohort design, court substantiated cases of childhood physical and sexual abuse and neglect during 1961 - 1971 were matched with nonabused and nonneglected children in one Midwestern metropolitan county area and followed into adulthood (mean age 41). Data contains demographics, family background, education and other cognitive tests, psychiatric tests, and economic status and productivity in 2003-2004.</td>
<td>807 individuals who were part of the prospective cohort design and had non-missing outcome measures in 2003-2004.</td>
<td>Child maltreatment reduces earnings in 2003-2004 by approximately $5000.</td>
</tr>
</tbody>
</table>
B.9. Appendix Figures

Figure B.1.: Bin Scatter of Household Income with Health Indices
Figure B.2.: Bin Scatter of Household Income with Family Based Indices

- **Family Environment**: $b_{on} = 0.053$, $t = 1.06$, $p = 0.293$
- **Family Network**: $b_{on} = 0.036$, $t = 0.90$, $p = 0.739$
- **Relationship with Parents**: $b_{on} = -0.003$, $t = 0.00$, $p = 0.999$
- **Parenting**: $b_{on} = 0.001$, $t = 0.00$, $p = 0.999$
Figure B.3.: Bin Scatter of Household Income with Childhood Experiences Indices
Figure B.4.: Bin Scatter of Household Income with Neighborhood Indices
Figure B.5.: Bin Scatter of Household Income with Education

\[ b \text{ on } X = 0.411, \quad b \text{ on } X^2 = 0.015 \]
\[ p\text{-value for coefficients jointly equal to 0} = 0.000 \]
Figure B.6.: Correlations With Parental Income Controls, by Race
Figure B.7.: Correlations With Parental Income Controls, by Gender
Figure B.8: Correlations With Parent Inc. Controls, by Parental Income
Figure B.9.: Non-Parametric, by Race
Figure B.10: Non-Parametric, by Gender
<table>
<thead>
<tr>
<th><strong>Figure B.11.</strong> Non-Parametric, by Parental Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Household Income</td>
</tr>
<tr>
<td>*Education (+)</td>
</tr>
<tr>
<td>*IPIP index (+)</td>
</tr>
<tr>
<td>*Ever in Trouble with Police Before 18 (−)</td>
</tr>
<tr>
<td>*Resilience (−)</td>
</tr>
<tr>
<td>*Parental Income Index (+)</td>
</tr>
<tr>
<td>*Self Esteem Index (+)</td>
</tr>
<tr>
<td>*Number of Adult Relationships Trusted (+)</td>
</tr>
<tr>
<td>*Growth Mindset (+)</td>
</tr>
<tr>
<td>0 1 2 3 4 5</td>
</tr>
<tr>
<td>Poor  Avg</td>
</tr>
<tr>
<td>(b) Adjusted Income</td>
</tr>
<tr>
<td>*Self Esteem Index (+)</td>
</tr>
<tr>
<td>*Number of Adult Relationships Trusted (+)</td>
</tr>
<tr>
<td>*Beliefs about Success (−)</td>
</tr>
<tr>
<td>Education (+)</td>
</tr>
<tr>
<td>0 2 4</td>
</tr>
<tr>
<td>Poor  Avg</td>
</tr>
<tr>
<td>(c) Individual Income</td>
</tr>
<tr>
<td>*Self Esteem Index (+)</td>
</tr>
<tr>
<td>*Relationship with Parents (−)</td>
</tr>
<tr>
<td>*Lived with Mother/Stepmother (−)</td>
</tr>
<tr>
<td>*Parents/Self Drink/Use Drugs Before 18 (−)</td>
</tr>
<tr>
<td>0 2 4</td>
</tr>
<tr>
<td>Poor  Avg</td>
</tr>
<tr>
<td>(d) Adult Mental Illness</td>
</tr>
<tr>
<td>*Self Esteem Index (+)</td>
</tr>
<tr>
<td>*Education (+)</td>
</tr>
<tr>
<td>*IPIP index (+)</td>
</tr>
<tr>
<td>*Ever in Trouble with Police Before 18 (−)</td>
</tr>
<tr>
<td>*Risk Attitudes as Teenager (−)</td>
</tr>
<tr>
<td>*Resilience (−)</td>
</tr>
<tr>
<td>*Parental Income Index (+)</td>
</tr>
<tr>
<td>*Number of Adult Relationships Trusted (+)</td>
</tr>
<tr>
<td>0 2 4</td>
</tr>
<tr>
<td>Poor  Avg</td>
</tr>
<tr>
<td>(e) Drug and Alcohol Use</td>
</tr>
<tr>
<td>*Lived with Mother/Stepmother (−)</td>
</tr>
<tr>
<td>*Relationship with Parents (−)</td>
</tr>
<tr>
<td>*Physical Illness before 16 (−)</td>
</tr>
<tr>
<td>*Mental Illness Before 16 (−)</td>
</tr>
<tr>
<td>0 2 4</td>
</tr>
<tr>
<td>Poor  Avg</td>
</tr>
<tr>
<td>(f) Adult Physical Health</td>
</tr>
</tbody>
</table>

346
Figure B.12.: Correlation Between Income and Locus of Control
Figure B.13.: Non-Parametric Correlations, Dobbie and Fryer (2013)
B.10. Appendix Tables
<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>Adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NLSY (1)</td>
<td>NLSY (2)</td>
</tr>
<tr>
<td>Top-coded Individual Income</td>
<td>28,311.881</td>
<td>27,751.124</td>
</tr>
<tr>
<td>Age</td>
<td>48.689</td>
<td>52.579</td>
</tr>
<tr>
<td>Female</td>
<td>0.527</td>
<td>0.513</td>
</tr>
<tr>
<td>Black</td>
<td>0.432</td>
<td>0.384</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.096</td>
<td>0.145</td>
</tr>
<tr>
<td>Poor appetite or Overeating</td>
<td>1.859</td>
<td>1.520</td>
</tr>
<tr>
<td>Trouble concentrating on things</td>
<td>1.562</td>
<td>1.588</td>
</tr>
<tr>
<td>Feeling depressed</td>
<td>1.656</td>
<td>1.514</td>
</tr>
<tr>
<td>Trouble falling asleep</td>
<td>2.184</td>
<td>1.997</td>
</tr>
</tbody>
</table>

Panel A: Demographics

Panel B: Adult Mental Illness Sub-Categories

Panel C: Psychological Traits

Table B.4.: Summary Statistics by Survey

Notes: This table presents summary statistics for different survey samples. Column (1) presents weighted averages for the authors’ survey in Shelby County, TN, Tulsa County, OK, and Jefferson and Orleans Parishes, LA. Columns (2) and (3) present weighted averages for respondents in the National Longitudinal Survey of Youth (NLSY) 1979 who were classified as poor in 1979. Column (4) presents the p-value for the difference in values between columns (1) and (2). Column (5) presents the p-value for the difference in values between columns (1) and (3). Weights for column (1) are sampling weights taken from the survey. Weights for column (2) are sampling weights taken from the 2014 wave of NLSY’79. Weights for column (3) were calculated to age-gender-race adjust the NLSY sample to look like the survey sample. Age groups are (a) 49-51 (b) 52-54 and (c) 55-57. Gender groups are (a) male and (b) female. Race groups are (a) Black, (b) Hispanic and (c) Other. Individual income for the NLSY sample was taken from its most recent 2014 wave. Individual incomes from the survey were top coded using the same rules as the NLSY sample. Age, gender and race information for the NLSY sample was taken from its 1978 screener information. For adult mental illness sub-categories, NLSY respondents were asked these questions once they turned 40 and then once when they turned 50. We use the response from when they turned 50. If that was missing and there was a non-missing response from when they turned 40, we used their previous response. All adult mental illness responses were categorical where 1 = Rarely/None of the time/1 Day, 2 = Some/A little of the time/1-2 Days, 3 = Occasionally/Moderate amount of the time/3-4 Days, and 4 = Most/All of the time/5-7 Days. For Rotter Locus of Control, we use NLSY’s 2014 measure. The Rotter score for our survey and the NLSY sample is a score which is between 4-16 where a higher score indicates more internal locus of control. For Rosenberg Self Esteem, we use NLSY’s 2006 measure. Rosenberg self esteem score for our survey and the NLSY sample is a score between 0-30 where a higher score indicates higher self esteem. All variables are explained in detail in the Data Appendix.
### Table B.5.: Summary Statistics (Survey vs. National)

<table>
<thead>
<tr>
<th></th>
<th>Survey</th>
<th>Census</th>
<th><em>p</em>-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>(1)</strong></td>
<td><strong>(2)</strong></td>
<td><strong>(3)</strong></td>
</tr>
<tr>
<td><strong>Panel A: Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household Income</td>
<td>47,483.895</td>
<td>91,850.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Female</td>
<td>0.527</td>
<td>0.508</td>
<td>0.369</td>
</tr>
<tr>
<td>Black</td>
<td>0.437</td>
<td>0.127</td>
<td>0.000</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.097</td>
<td>0.178</td>
<td>0.000</td>
</tr>
<tr>
<td>High School/GED/Incomplete College</td>
<td>0.488</td>
<td>0.407</td>
<td>0.000</td>
</tr>
<tr>
<td>Two-year Associate Degree</td>
<td>0.119</td>
<td>0.064</td>
<td>0.000</td>
</tr>
<tr>
<td>Bachelor or Some Post-Graduate Degree</td>
<td>0.208</td>
<td>0.231</td>
<td>0.170</td>
</tr>
<tr>
<td>Observations</td>
<td>900</td>
<td>3,156,487</td>
<td></td>
</tr>
</tbody>
</table>

Notes: This table presents summary statistics for different survey samples. Column (1) presents weighted averages for the authors' survey in Shelby County, TN, Tulsa County, OK, and Jefferson and Orleans Parishes, LA. Column (2) presents weighted averages for participants in the American Community Survey 2016 (ACS 2016). Column (3) presents the *p*-value for the difference in values between columns (1) and (2). Weights for column (1) are sampling weights taken from the survey. Weights for column (2) are sampling weights taken from ACS 2016. All variables are explained in detail in the Data Appendix.
<table>
<thead>
<tr>
<th>Panel A: Demographics</th>
<th>Phone</th>
<th>Non-Eligible</th>
<th>Eligible</th>
<th>Agreed to</th>
<th>Paper</th>
<th>p-value</th>
<th>p-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poor Now?</td>
<td>0.212</td>
<td>0.095</td>
<td>0.385</td>
<td>0.419</td>
<td>0.413</td>
<td>0.000</td>
<td>0.001</td>
<td>0.029</td>
</tr>
<tr>
<td>Age</td>
<td>52.982</td>
<td>53.522</td>
<td>52.191</td>
<td>50.865</td>
<td>50.990</td>
<td>0.005</td>
<td>0.001</td>
<td>0.012</td>
</tr>
<tr>
<td>Female</td>
<td>0.553</td>
<td>0.571</td>
<td>0.527</td>
<td>0.526</td>
<td>0.545</td>
<td>0.000</td>
<td>0.932</td>
<td>0.155</td>
</tr>
<tr>
<td>Panel B: Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.472</td>
<td>0.572</td>
<td>0.326</td>
<td>0.310</td>
<td>0.310</td>
<td>0.000</td>
<td>0.094</td>
<td>0.186</td>
</tr>
<tr>
<td>Black</td>
<td>0.226</td>
<td>0.198</td>
<td>0.266</td>
<td>0.334</td>
<td>0.357</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.251</td>
<td>0.181</td>
<td>0.355</td>
<td>0.291</td>
<td>0.271</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Panel C: Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than HS</td>
<td>0.073</td>
<td>0.023</td>
<td>0.145</td>
<td>0.120</td>
<td>0.102</td>
<td>0.000</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>Some HS, Incomplete</td>
<td>0.074</td>
<td>0.051</td>
<td>0.108</td>
<td>0.109</td>
<td>0.105</td>
<td>0.000</td>
<td>0.926</td>
<td>0.671</td>
</tr>
<tr>
<td>HS/GED/Incomplete Col.</td>
<td>0.396</td>
<td>0.374</td>
<td>0.427</td>
<td>0.434</td>
<td>0.424</td>
<td>0.000</td>
<td>0.478</td>
<td>0.820</td>
</tr>
<tr>
<td>Two Year Associate</td>
<td>0.099</td>
<td>0.103</td>
<td>0.092</td>
<td>0.103</td>
<td>0.117</td>
<td>0.137</td>
<td>0.068</td>
<td>0.002</td>
</tr>
<tr>
<td>Bachelor or Graduate</td>
<td>0.183</td>
<td>0.224</td>
<td>0.124</td>
<td>0.124</td>
<td>0.134</td>
<td>0.000</td>
<td>0.914</td>
<td>0.243</td>
</tr>
</tbody>
</table>

Observations 6,459 3,842 2,617 1,227 928

Table B.6.: Summary Statistics by Sample

Notes: This table presents summary statistics for different samples. Column (1) presents averages for all participants with completed phone screeners. Column (2) presents averages for all participants who were deemed non-eligible to participate in the survey. Column (3) presents averages for all participants who were deemed eligible to participate in the survey. Column (4) presents averages for all participants who were deemed eligible and who agreed to participate in the paper survey. Column (5) presents averages for all participants who participated in the paper survey. Column (6) presents the p-value for the difference in values between people who were deemed non-eligible and people who were deemed eligible for the paper survey. Column (7) presents the p-value for the difference in values between people who were deemed eligible but did not agree and people who were deemed eligible and agreed to participate in the survey. Column (8) presents the p-value for the difference in values between people who finally participated in the paper survey and people who were deemed eligible but did not participate in the paper survey.
Table B.7: Correlation Matrix

The table presents pairwise correlations between all pairs of variables. The variables considered are: (1) Education, (2) Mental Illness Before 16, (3) Physical Illness Before 16, (4) Psychological Index, (5) Diet, (6) Family Environment, (7) Family Network, (8) Relationship with Parents, (9) Parenting, (10) Trust any Adults in Childhood, (11) Number of Adult Relationships Trusted, (12) Quality of Adult Relationships Trusted, (13) Adverse Childhood Experience, (14) Risky Attitudes as Teenager, (15) Trauma Before 18, (16) Lifetime Trauma, (17) Beliefs about Success, (18) Prob. of bottom 25 in top 20 percentile, (19) Frac. with fathers present, (20) Neighborhood Safety Index. Note that indices for lifetime trauma before 16 years of age and any lifetime trauma and the psychological index were created from sub-indices for these matrices only. Lifetime trauma before 16 Index was created out of 4 individual questions. Any lifetime trauma was created using 9 individual questions. Psychological index was created as a sum of 7 sub-indices – grit, resilience, IPIP, self esteem, self control, locus of control, and growth mindset. All variables are explained in detail in the Appendix B.3.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>0.06</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>-0.01</td>
<td>0.32</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>0.31</td>
<td>-0.23</td>
<td>0.12</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5)</td>
<td>0.03</td>
<td>-0.13</td>
<td>-0.18</td>
<td>0.10</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6)</td>
<td>-0.05</td>
<td>-0.29</td>
<td>-0.14</td>
<td>0.21</td>
<td>0.22</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(7)</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.05</td>
<td>0.16</td>
<td>0.11</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(8)</td>
<td>-0.03</td>
<td>-0.28</td>
<td>-0.14</td>
<td>0.10</td>
<td>0.17</td>
<td>0.30</td>
<td>0.10</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>-0.01</td>
<td>-0.28</td>
<td>0.17</td>
<td>0.17</td>
<td>0.38</td>
<td>0.48</td>
<td>0.16</td>
<td>0.51</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>0.04</td>
<td>-0.16</td>
<td>-0.07</td>
<td>0.13</td>
<td>0.17</td>
<td>0.22</td>
<td>0.11</td>
<td>0.22</td>
<td>0.29</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(11)</td>
<td>0.10</td>
<td>-0.13</td>
<td>0.00</td>
<td>0.13</td>
<td>0.16</td>
<td>0.18</td>
<td>0.09</td>
<td>0.09</td>
<td>0.24</td>
<td>0.37</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(12)</td>
<td>0.12</td>
<td>-0.13</td>
<td>0.03</td>
<td>0.13</td>
<td>0.09</td>
<td>0.20</td>
<td>0.04</td>
<td>0.24</td>
<td>0.31</td>
<td>0.14</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(13)</td>
<td>-0.00</td>
<td>0.42</td>
<td>0.31</td>
<td>-0.15</td>
<td>-0.32</td>
<td>-0.43</td>
<td>-0.10</td>
<td>-0.46</td>
<td>-0.57</td>
<td>-0.21</td>
<td>-0.14</td>
<td>-0.24</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(14)</td>
<td>-0.06</td>
<td>0.44</td>
<td>0.32</td>
<td>-0.33</td>
<td>-0.08</td>
<td>-0.31</td>
<td>0.04</td>
<td>-0.20</td>
<td>-0.34</td>
<td>-0.05</td>
<td>-0.03</td>
<td>-0.08</td>
<td>0.36</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(15)</td>
<td>-0.11</td>
<td>0.31</td>
<td>0.32</td>
<td>-0.09</td>
<td>-0.13</td>
<td>-0.27</td>
<td>-0.05</td>
<td>-0.26</td>
<td>-0.33</td>
<td>-0.07</td>
<td>-0.07</td>
<td>-0.13</td>
<td>0.51</td>
<td>0.41</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(16)</td>
<td>-0.12</td>
<td>0.23</td>
<td>0.28</td>
<td>-0.16</td>
<td>-0.11</td>
<td>-0.13</td>
<td>0.04</td>
<td>-0.12</td>
<td>-0.16</td>
<td>-0.06</td>
<td>-0.04</td>
<td>-0.08</td>
<td>0.34</td>
<td>0.30</td>
<td>0.32</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(17)</td>
<td>0.01</td>
<td>-0.05</td>
<td>-0.01</td>
<td>0.154</td>
<td>0.03</td>
<td>0.09</td>
<td>0.04</td>
<td>0.08</td>
<td>0.06</td>
<td>0.07</td>
<td>0.04</td>
<td>0.01</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>-0.06</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(18)</td>
<td>0.18</td>
<td>-0.03</td>
<td>0.01</td>
<td>0.11</td>
<td>0.08</td>
<td>-0.00</td>
<td>-0.01</td>
<td>-0.00</td>
<td>-0.00</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>-0.04</td>
<td>-0.03</td>
<td>-0.04</td>
<td>-0.01</td>
<td>-0.05</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(19)</td>
<td>0.14</td>
<td>-0.00</td>
<td>0.01</td>
<td>0.14</td>
<td>0.03</td>
<td>0.003</td>
<td>0.03</td>
<td>0.01</td>
<td>0.04</td>
<td>0.03</td>
<td>0.07</td>
<td>0.09</td>
<td>-0.00</td>
<td>-0.08</td>
<td>0.00</td>
<td>-0.01</td>
<td>-0.06</td>
<td>0.77</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>(20)</td>
<td>0.00</td>
<td>-0.22</td>
<td>-0.14</td>
<td>0.12</td>
<td>0.17</td>
<td>0.232</td>
<td>0.06</td>
<td>0.16</td>
<td>0.22</td>
<td>0.12</td>
<td>0.10</td>
<td>0.03</td>
<td>-0.31</td>
<td>-0.31</td>
<td>-0.22</td>
<td>-0.15</td>
<td>-0.08</td>
<td>0.10</td>
<td>0.15</td>
<td>1.00</td>
</tr>
</tbody>
</table>
C. Appendix to Chapter 3

C.1. Decomposing By More Categories

Similar to variance decomposition, skewness decomposition can also be easily extended to accommodate with linear models. Assume the following simple linear model when $Y$ is standardized:

$$ Y = \sum_i X_i $$

Using simple algebra

$$ \mu_3(Y) = \sum_i \mu_3(X_i) + \sum_i \sum_{j \neq i} COV(X_i^2, X_j) + \sum_i \sum_{j \neq i} \sum_{k \neq i,j} E[X_iX_jX_k] $$  (C.1)

Therefore, we can decompose the skewness of $Y$ into a linear combination of the skewness of its linear components, the covariance of their second and first moments, and the triple multiplication of all three distinguished components. I will call this decomposition using Equation C.1 - linear skewness decomposition. Though this decomposition includes a lot of different terms, many of them has an expectation of zero.

For example, writing $Y$ as the sum of its conditional expectation in $X$ and a residual $\varepsilon$

$$ Y = E[Y|X] + \varepsilon $$
and using Equation C.1 yields Equation 3.4, using the law of iterated expectations over X.

Alternatively we can estimate any linear model, and use this formula. This is useful to compare occupations directly to other categories. I show this for the equation

$$\ln w_i = occ_i + ind_i + \varepsilon_i$$

where $occ_i$ and $ind_i$ are occupation and industry dummies. I then decompose the increase in skewness by Equation C.1. Figure C.4 presents the results. Most of the increase in skewness comes from the correlation of $occ_i$ and $\varepsilon_i^2$. That’s the correlation of the part of occupation premium that is orthogonal to industry, and inequality within occupation and industry. The equivalent component for industries (in green) is negligible. All other components, such as the skewness between occupations or industries, the correlation of occupation premium and industry premium variance and others are aggregated and plotted in red. All together they comprise only a small share of the increase.

To do the same exercise for occupations with observable skills I estimate a Mincer equation with occupational dummies

$$\ln w_i = occ_i + \beta X_i + \varepsilon_i$$

Similar to the case with industries, the covariance for occupations and residuals is still capturing almost half of the increase. Note that this is the correlation with the occupation premium, for workers with the same level of skills.

However, as opposed to the case of industries, the correlation of skill group mean wage, and variance

$$COV \left( \beta X_i, \varepsilon_i^2 \right)$$
also seem to matter. To further investigate that I decompose $\beta X_i$ to mean occupation skill and within occupation skill difference

$$E [\beta X_i | occ_i] + (\beta X_i - E [\beta X_i | occ_i])$$

Decomposing by the four components

$$\ln w_i = occ_i + E [\beta X_i | occ_i] + (\beta X_i - E [\beta X_i | occ_i]) + \varepsilon_i$$

I find that all of the increase in $COV (\beta X_i, \varepsilon_i^2)$ comes from $COV (E [\beta X_i | occ_i], \varepsilon_i^2)$ as can be shown in Figure C.5. At total, the main two components are the correlation of $\varepsilon^2$ with both $occ_i$ and $E [\beta X_i | occ_i]$. This means that the correlation of variance with occupational wage levels is stemming from both, occupation premium ($occ_i$) and mean skill level at the occupation ($E [\beta X_i | occ_i]$). But similar to industries, categories that are unrelated to occupations are still negligible.

C.2. Appendix Figures
**Figure C.1.**: Skewness Decomposition by 3-Digit Occupation With Imputed Wages

Changes since base year (1992). Wages at the top and bottom 5% were dropped (see Section 3.4). Includes wages that were imputed.

Source: CPS Outgoing Rotation Groups

**Figure C.2.**: Skewness Decomposition by 3-Digit Industry

Changes since base year (1992). Wages at the top and bottom 5% were dropped (see Section 3.4).

Source: CPS Outgoing Rotation Groups
Figure C.3.: Skewness Decomposition by Education and Experience
Changes since base year (1992). Wages at the top and bottom 5% were dropped (see Section 3.4).
Source: CPS Outgoing Rotation Groups

Figure C.4.: Skewness Decomposition by Occupation and Industry
Changes since base year (1992). Wages at the top and bottom 5% were dropped (see Section 3.4). Decomposing based on Equation C.2. I plot only two components separately and the rest are aggregated (in red). $COV(occ, \varepsilon^2)$ and $COV(ind, \varepsilon^2)$ are the covariance of occupation and industry premium with the unexplained variance.
Source: CPS outgoing rotation groups.
Figure C.5.: Skewness Decomposition by Occupation, School, Experience

Changes since base year (1992). Wages at the top and bottom 5% were dropped (see Section 3.4). Decomposing based on Equation C.2. I plot only two components separately and the rest are aggregated (in red).

Source: CPS outgoing rotation groups.
**Figure C.6.: Occupation Premium by Skill Percentile with Instruments**

Difference in log predicted wage for workers in routine versus abstract/manual occupation for three percentiles at the distribution of $\theta_i$. Estimated with an interactive fixed effect model, using years of education as instruments. $\theta_i$ are defined net of age and cohort. Routine workers are defined as workers in administrative, production or operator occupations, classified by the first occupational coding digit. Similarly manual includes all service, sales and agriculture workers.

Source: PSID