The Evolution of Labor Earnings Risk in the U.S. Economy*

Flavio Cunha James Heckman

First draft, May 2005

This draft, February 26, 2007

Abstract

A large empirical literature documents a rise in wage inequality in the American economy. It is silent on whether the increase in inequality is due to greater heterogeneity in the components of earnings that are predictable by agents or whether it is due to greater uncertainty faced by agents. Using choice data combined with earnings data, we find that both predictable and unpredictable components have increased in recent years, and that the increase in uncertainty is greater for unskilled workers. For both groups, roughly 70% of the increase in wage variability within schooling groups is due to uncertainty. Roughly 20% of the increase in the variance of returns to schooling is due to increased uncertainty.

JEL codes: D3; J8
Key words: wage inequality, uncertainty, sorting

Flavio Cunha
Department of Economics
University of Chicago
1126 E. 59th St.
Chicago, IL 60637
Phone: (773) 256-6141
Fax: (773) 256-6313
E-mail: flavio@uchicago.edu

James J. Heckman
Department of Economics
University of Chicago
1126 E. 59th Street
Chicago IL 60637
Phone: (773) 702-0634
Fax: (773) 702-8490
E-mail: jjh@uchicago.edu

*This research was supported by NIH R01-HD-043411 and NSF SES-0241858. Cunha is grateful to the Claudio Haddad Dissertation Fund at the University of Chicago for research support. This research is an outgrowth of research reported in Cunha, Heckman, and Navarro (2005). We are grateful to Ray Fair, Lars Hansen, Pat Kehoe, Robert Lucas, Salvador Navarro, Tom Sargent, Robert Shimer, Robert Townsend and Kenneth Wolpin for comments on various drafts. This version has benefited from comments received at the Money and Banking Workshop, University of Chicago, November 21, 2006. We have also presented this paper in the Ely Lectures at Johns Hopkins University, April 2005, the 9th Econometric Society World Congress at University College London, August 2005, the Economic Dynamics Working Group at University of Chicago, October 2005, the Empirical Dynamic General Equilibrium Conference at the Centre for Applied Microeconometrics, December 2005, Macroeconomics of Imperfect Risk Sharing Conference at the University of California at Santa Barbara, May 2006, the 2006 Meetings of the Society for Economic Dynamics, July 2006, as part of the Koopmans Memorial Lectures at Yale, September 2006, the Federal Reserve Bank of Minneapolis Applied Micro Workshop, October 2006, and Tom Sargent's Macro Reading Group at New York University, October 2006. The website for this paper is http://jenni.uchicago.edu/evo-earn/.

1 Introduction

A large literature documents an increase in wage inequality in the American economy over the 1970's and 1980's (see, for example, Levy and Murnane, 1992, or Katz and Autor, 1999). This increase in wage inequality has occurred both within and between education-experience groups.

Increased variability in wages across people over time is not the same as increased uncertainty in wages. This paper estimates how much of the recent increase in wage inequality is due to an increase in heterogeneity that is predictable by the agents at the age they make their college attendance decisions but is not known to the observing economist, and how much is due to uncertainty.

We demonstrate that an increase in microeconomic uncertainty plays an important role in explaining the recent increase in wage inequality. Our findings are consistent with the analysis of Gottschalk and Moffitt (1994), who document an increase in "earnings instability" (the $\varepsilon_{s,t}$), demonstrating that the variance of transitory components rose considerably from the period 1970–1978 to the period 1979–1987. However, their framework cannot distinguish uncertainty from variability. Transitory components as measured by a statistical decomposition of earnings may be perfectly predictable by agents or totally unpredictable. This paper uses schooling choices to estimate the information sets of agents at the age college enrollment decisions are made. We show that unforecastable components in labor income have increased across cohorts. Earnings instability, or turbulence, has increased substantially.[1]

We model schooling and earnings equations jointly. Modelling schooling choices is more than an econometric exercise to correct for selection bias in earnings although it has that benefit. Schooling choices are a source of information that allows us to separate what is known and acted on by individuals at the time schooling choices are made — what we call heterogeneity — from what is not known — what we call uncertainty.

The method we use to measure the increase in wage uncertainty in the recent American labor market is based on the following simple idea. Suppose that we have data on decisions about a choice variable S. The choice variable is assumed to depend, in part, on current and future income, $Y_1, Y_2, \ldots, Y_T$, where T is the horizon for agent decision making, through its present value: $PV = \sum_{t=1}^{T} \left(Y_t / \left(1+\rho\right)^{t-1} \right)$, where ρ is the discount rate.

[1] See Ljungqvist and Sargent (2004), who discuss the recent rise of turbulence in the economy.

In the first period, agents only imperfectly predict their future earnings using information $\mathcal{I}$. Thus, S depends on future income, $Y_1, \ldots, Y_T$, through $E(PV \mid \mathcal{I})$, where "E" denotes expectation. If, after the choice is made, we actually observe $Y_1, \ldots, Y_T$, we can construct PV *ex post*. If the information set is properly specified, the residual corresponding to the component of PV that is not forecastable in the first period, $V = PV - E(PV \mid \mathcal{I})$, should not predict S. $E(PV \mid \mathcal{I})$ is predictable heterogeneity, allowing for information heterogeneity among agents. V is a measure of uncertainty.[2]

This paper develops and applies a method for inferring $\mathcal{I}$ from panel data where the choice is college going. Agents have two potential income streams corresponding to the earnings associated with going to college and the earnings associated with not going to college. Because we observe the earnings streams of individuals in only one of two possible states (college / no college), it is necessary to account for the missing counterfactual earnings of each person in order to measure unpredictable components. This is why we worry about self selection problems in this paper.

The rest of this paper is in five parts. Part 2 presents the model. Part 3 presents the econometrics and the empirical results. Part 4 discusses a more general framework. Part 5 concludes.

2 The Model

We estimate the information sets of the agents. We identify the agent information sets by analyzing both the choices and the outcomes associated with choices made by the agents.

2.1 Earnings Equations

To motivate our econometric procedures, we start by describing the earnings equations for $t = 1, \ldots, T$, which are life cycle outcomes over horizon T. We assume that $(Y_{0,t}, Y_{1,t})$, $t = 1, \ldots, T$, have finite means and can be expressed in terms of conditioning variables X in the following manner:

$$Y_{0,t} = X\beta_{0,t} + U_{0,t} \tag{1}$$

$$Y_{1,t} = X\beta_{1,t} + U_{1,t}, \qquad t = 1, \ldots, T. \tag{2}$$

[2] The Sims (1972) test for noncausality is based on this idea. Whereas he tests whether future Y predicts current S, we measure what fraction of future Y predicts current S.

The error terms $U_{s,t}$ are assumed to satisfy $E\left(U_{s,t} \mid X\right) = 0$, $s = 0, 1$.

2.2 Choice Equations

We assume that agents make schooling choices based on expected present value income maximization given information set $\mathcal{I}$. Write the index I of present values as

$$I = E\left[\sum_{t=1}^{T}\left(\frac{1}{1+\rho}\right)^{t-1}\left(Y_{1,t} - Y_{0,t}\right) - C\middle|\mathcal{I}\right], \tag{3}$$

where C is the cost of attending college. We denote by Z and U_C the observable and unobservable determinants of costs, respectively. We assume that costs can be written as

$$C = Z\gamma + U_C. \tag{4}$$

If we define $\mu_I(X, Z) = \sum_{t=1}^{T}\left(\frac{1}{1+\rho}\right)^{t-1} X\left(\beta_{1,t} - \beta_{0,t}\right) - Z\gamma$ and $U_I = \sum_{t=1}^{T}\left(\frac{1}{1+\rho}\right)^{t-1}\left(U_{1,t} - U_{0,t}\right) - U_C$, and substitute (1), (2), and (4) into (3) we obtain

$$I = E\left[\mu_I(X, Z) + U_I \middle| \mathcal{I}\right]. \tag{5}$$

U_I is the error term in the choice equation and it may or may not include $U_{1,t}$, $U_{0,t}$, or U_C, depending on what is in the agent's information set. Similarly, $\mu_I(X, Z)$ may only be based on expectations of future X and Z at the time schooling decisions are made. Schooling is generated by

$$S = \mathbf{1}\left[I \geq 0\right]. \tag{6}$$

2.3 Test Score Equations

Aside from data on earnings and choices, we also have data on a set of cognitive test score equations. Let M_k denote the agent's score on the k^{th} test. Assume that the M_k have finite means and can be expressed in terms of conditioning variables X^M. Write

$$M_k = X^M\beta_k^M + U_k^M, \qquad k = 1, 2, \ldots, K. \tag{7}$$

The test equations are introduced here because we expect both the decision to attend college and realized earnings to depend on the cognitive skills that the agent has at the time schooling choices are made. Test scores facilitate but are not essential to our identification strategy.

2.4 Heterogeneity and Uncertainty

To focus on main ideas, assume that $X \in \mathcal{I}$. Note that we can always write the earnings of school level s at age t as

$$Y_{s,t} = X\beta_{s,t} + E\left(U_{s,t} \mid \mathcal{I}\right) + \left[U_{s,t} - E\left(U_{s,t} \mid \mathcal{I}\right)\right].$$

The component $E\left(U_{s,t} \mid \mathcal{I}\right)$ is available to the agent to help make schooling choices. It affects realized earnings. The component $U_{s,t} - E\left(U_{s,t} \mid \mathcal{I}\right)$ does not enter the schooling equation because it is unknown at the time schooling decisions are made. However, it affects realized earnings.

To determine the unobservable components that are in the information set of the agent we need to determine which specification of the information set $\mathcal{I}$ best characterizes the dependence between schooling choices and future earnings. We can determine the components that are not in the information set of the agent by varying the specification of $\mathcal{I}$ in order to obtain the best possible fit of the distribution of $Y_{s,t}$ and schooling choices. In the next section we describe how we use factor models to represent both $E\left[U_{s,t}|\mathcal{I}\right]$ and $\left(U_{s,t} - E\left[U_{s,t}|\mathcal{I}\right]\right)$ in a framework that is convenient for testing and estimation.

2.5 Factor Models

To demonstrate our approach to determining the elements in the information set of the agent, we start by considering the test score equations. We break the error term U_k^M in the test score equations into two components. The first component is a factor, θ_1, that is common across all test score equations. The second component is unique to test score equation k, ε_k^M. In this notation, we can write equation (7) as

$$M_k = X^M \beta_k^M + \alpha_k^M \theta_1 + \varepsilon_k^M. \tag{8}$$

Following the psychometric literature, the factor θ_1 is a latent cognitive ability which potentially affects all test scores. We assume that θ_1 is independent of X^M and ε_k^M. The ε_k^M are mutually

independent and independent of θ_1. Modelling test scores in this fashion allows them to be noisy measures of cognitive ability.

2.5.1 Earnings and Choice Equations

We decompose the error terms in the earnings equations into three components. The first component is the cognitive factor θ_1. The second component is a "productivity" factor θ_2 which affects earnings and schooling choices, but not test scores. In our empirical work, we fit models with as many as six factors, but for expositional purposes, in this section we work with a two factor model. The third component of the earnings error term is the idiosyncratic error term which affects only the period-t, schooling-s earnings equation, $\varepsilon_{s,t}$. We assume that $U_{0,t}$ and $U_{1,t}$ can be written in factor-structure form

$$U_{i,t} = \alpha_{1,i,t}\theta_1 + \alpha_{2,i,t}\theta_2 + \varepsilon_{i,t}\,, \qquad i = 0, 1,$$

so that (1) and (2) can be written as

$$Y_{0,t} = X\beta_{0,t} + \alpha_{1,0,t}\theta_1 + \alpha_{2,0,t}\theta_2 + \varepsilon_{0,t} \tag{9}$$

and

$$Y_{1,t} = X\beta_{1,t} + \alpha_{1,1,t}\theta_1 + \alpha_{2,1,t}\theta_2 + \varepsilon_{1,t}. \tag{10}$$

We assume that factor θ_j is independent from X, $\varepsilon_{s,t}$, and θ_l for $l \neq j$ and for all s, t. The $\varepsilon_{\ell,t}$, $\ell = 0, 1$ and $t = 1, \ldots, T$, are mutually independent. The cost equation is decomposed like the earnings equations, so that (4) can be rewritten as

$$C = Z\gamma + \alpha_{1,C}\theta_1 + \alpha_{2,C}\theta_2 + \varepsilon_C. \tag{11}$$

Given the factor specifications in (9), (10), and (11), we can rewrite the schooling choice equation

as

$$I = E\left[\begin{array}{l} \sum_{t=1}^{T} \left(\frac{1}{1+\rho}\right)_i^{t-1} X\left(\beta_{1,t} - \beta_{0,t}\right) - Z\gamma + \theta_1 \left[\sum_{t=1}^{T} \left(\frac{1}{1+\rho}\right)^{t-1} \left(\alpha_{1,1,t} - \alpha_{1,0,t}\right) - \alpha_{1,C}\right] \\ \quad +\theta_2 \left[\sum_{t=1}^{T} \left(\frac{1}{1+\rho}\right)^{t-1} \left(\alpha_{2,1,t} - \alpha_{2,0,t}\right) - \alpha_{2,C}\right] + \sum_{t=1}^{T} \left(\frac{1}{1+\rho}\right)^{t-1} \left(\varepsilon_{1,t} - \varepsilon_{0,t}\right) - \varepsilon_C \end{array} \middle| \mathcal{I} \right]. \tag{12}$$

We assume that for all distinct subscripts the ε's are mutually independent and independent of the X, Z, and (θ_1, θ_2).

2.6 The Estimation of the Information Set

We now show how to determine the information set $\mathcal{I}$ of the agent at the age schooling choices are made by exploiting the structure of factor models. Assume that X, Z, and ε_C are in the information set $\mathcal{I}$. To economize on notation, define

$$\alpha_{k,I} = \sum_{t=1}^{T} \left(\frac{1}{1+\rho}\right)^{t-1} \left(\alpha_{k,1,t} - \alpha_{k,0,t}\right) - \alpha_{k,C} \text{ for } k = 1, 2. \tag{13}$$

Suppose that it is claimed that $\{\theta_1, \theta_2\} \subset \mathcal{I}$, but $\varepsilon_{s,t} \notin \mathcal{I}$. Given the definitions of $\alpha_{1,I}, \alpha_{2,I}$ and $\mu_I(X, Z)$, if this hypothesis is true, the index governing schooling choices is

$$I = \mu_I(X, Z) + \alpha_{1,I}\theta_1 + \alpha_{2,I}\theta_2 + \varepsilon_C. \tag{14}$$

Suppose for the sake of argument that we know $\mu_I(X, Z)$ and $\beta_{s,t}$ for all s and t. From discrete choice analysis it is well established that under standard conditions, we can proceed for the purposes of model identification as if we know I up to scale.[3] Given data $Y_{1,1}$ on X and Z, we can identify the covariance between the terms $I - \mu_I(X, Z)$ and $Y_{1,1} - X\beta_{1,1}$. Under the hypothesis $\{\theta_1, \theta_2\} \subset \mathcal{I}$, this covariance is

$$\text{Cov}\left(I - \mu_I(X, Z), Y_{1,1} - X\beta_{1,1}\right) = \alpha_{1,I}\alpha_{1,1,1}\sigma^2_{\theta_1} + \alpha_{2,I}\alpha_{2,1,1}\sigma^2_{\theta_2}. \tag{15}$$

We can test the hypothesis $\{\theta_1, \theta_2\} \subset \mathcal{I}$ against many different alternative hypotheses. To fix ideas, consider the alternative hypothesis that proposes that $\theta_1 \in \mathcal{I}$, but $\theta_2 \notin \mathcal{I}$ and that

[3] See, e.g., Matzkin (1992).

$E[\theta_2 \mid \mathcal{I}] = 0$. If the alternative is valid, the expected present value of the gain from schooling (12) can be written as

$$I = \mu_I(X, Z) + \alpha_{1,I}\theta_1 + \varepsilon_C. \tag{16}$$

In this case, the covariance between the terms $I - \mu_I(X, Z)$ and $Y_{1,1} - X\beta_{1,1}$ is

$$\text{Cov}\left(I - \mu_I(X, Z), Y_{1,1} - X\beta_{1,1})\right) = \alpha_{1,I}\alpha_{1,1,1}\sigma^2_{\theta_1}. \tag{17}$$

The difference between (15) and (17) is the term $\alpha_{2,I}\alpha_{2,1,1}\sigma^2_{\theta_2}$ arising from the assumed greater information in the model generating (15). We can characterize a variety of tests of alternative information structures by defining parameters Δ_{θ_1} and Δ_{θ_2} such that

$$\text{Cov}\left(I - \mu_I(X, Z), Y_{1,1} - \mu_1(X)\right) - \Delta_{\theta_1}\alpha_{1,I}\alpha_{1,1,1}\sigma^2_{\theta_1} - \Delta_{\theta_2}\alpha_{2,I}\alpha_{2,1,1}\sigma^2_{\theta_2} = 0.$$

Agents know and act on the information contained in factors 1 and 2, so that $\{\theta_1, \theta_2\} \subset \mathcal{I}$, if we reject the hypothesis that both $\Delta_{\theta_1} = 0$ and $\Delta_{\theta_2} = 0$.

It remains to be shown that we can actually identify all of the parameters of the model, in particular, the function $\mu_I(X, Z)$, the parameters β and α in the test and earnings equations, the parameters of the cost functions, the distribution of the factors, F_θ, as well as the distribution of idiosyncratic components F_ε in the test, earnings and cost equations. Carneiro, Hansen, and Heckman (2003) present formal proofs of semi-parametric identification of this model. Appendix A presents an intuitive explanation of identification assuming normality of the unobservables. Normality is *not* required to secure identification, and our estimates are *not* based on normality assumptions.[4]

[4]There are at least two interpretations of the factor structure as we use it here. One is as a Gorman-Lancaster model of earnings as used by Heckman and Scheinkman (1987). The "α" are economy-wide prices known to the agents and the "θ" are endowments of the agent (e.g., abilities). Under this interpretation, agents learn about their abilities or productivities as encapsulated in θ. A second interpretation, and the one we prefer, is that agents may not know either all of their productivities or future prices at the time they make their schooling decisions. In this case, we write $U_{s,t} = \nu'_{s,t}\tau + \varepsilon_{s,t}$, $s = 0, 1$, where $\nu'_{s,t}$ are prices in sector s, and τ are quantities. Information updating on the *product* of $\nu_{s,t}$ and τ can produce a factor structure like that used in equations (9) and (10), but the factor loadings are not necessarily prices. For example, suppose that agents know scalar τ at the beginning of life. Price shocks arrive in sector s as random walk increments:

$$\nu_{s,1} = \nu_{s,0} + \phi_{s,1},$$

where $\phi_{s,1} \perp\!\!\!\perp \nu_{s,0}$, and agents know $\nu_{s,0}$. In the period "0", we have a one factor model, where $\alpha_{1,s,0} = \nu_{s,0}$. In the second period, we acquire a second component to the error term, $\phi_{s,1}\tau$, where $\phi_{s,1}$ is the random increment in prices. This innovation is common across persons as a result of the assumption competative labor markets. Thus, $\phi_{s,1}\tau$ is

3 Empirical Results

In order to study the evolution of the riskiness of labor earnings in the U.S. economy we analyze and compare two distinct samples. The first sample consists of white males born between 1957 and 1964. We obtain information on them from National Longitudinal Survey of Youth (NLSY/1979) data pooled from their birth cohort counterparts from the Panel Survey of Income Dynamics (PSID) data.[5] The second sample consists of white males born between 1941 and 1952 who are surveyed in the National Longitudinal Survey (NLS/1966) combined with their birth cohort counterparts from the PSID data.[6] We pool the surveys for comparable cohorts to increase sample sizes. In what follows, we refer to the samples as NLSY/1979 and NLS/1966, respectively. These data and our pooling methods are discussed in our data web appendix at http://jenni.uchicago.edu/evo-earn/.

We consider only two schooling choices: high school and college graduation. We use $s = 0$ to denote those who stop their schooling at high school and $s = 1$ to denote those who go to college. The Web Data Description Appendix Tables 1 and 2 present descriptive statistics of the NLS/1966 and NLSY/1979 samples, respectively. In both samples, college graduates have higher test scores, fewer siblings and parents with higher levels of education. In the NLSY/1979, college graduates are more likely to live in locations where the tuition for four-year college is lower. This is not true for the college graduates in NLS/1966.[7]

In our empirical analysis, we analyze labor income from ages 22 to 41. Web Data Appendix Tables 3 and 4 show mean and standard deviations for earnings in high school and college for NLSY/1979 and NLS/1966, respectively.[8] In both data sets, college graduates start off with lower mean labor income than high-school graduates.[9] The standard error of earnings increases with age for high school and college graduates in both data sets.

Both data sets have measures of cognitive test scores, which are the left-hand side variables in the

orthogonal to $\nu_{s,0}\tau$. In addition to incremental updating of prices, there may be multiple endowments $\tau_1, \tau_2, \ldots$ which are mutually independent and revealed to agents in different time periods. Thus, we can obtain a factor structure as a representation of a price-quantity updating process, and both prices and endowments can be revealed over time.

[5] See Miller (2004) for a description of the NLSY data and Hill, Duncan, and Marsden (1992) for a description of the PSID data.

[6] See documentation at http://www.nlsinfo.org/web-investigator/docs.php?mychrt=boys for a description of the NLS data.

[7] See Web Data Appendix for details on the construction of the tuition variables used in this paper.

[8] Earnings figures are adjusted for inflation using the CPI and we take the year 2000 as the base year.

[9] The overtaking age (the age when the mean earnings of college graduates equal the mean earnings of high school graduates) is 26 in both data sets. See Web Data Appendix Figures 9 and 10.

measurement system for cognitive ability (M in the notation of section 2). For the NLSY/1979 we use five components of the ASVAB test battery: arithmetic reasoning, word knowledge, paragraph comprehension, math knowledge and coding speed. We dedicate the first factor (θ_1) to this test system, and exclude other factors from it. This justifies our interpretation of θ_1 as ability.

In the NLS/1966, there are many different achievement tests, but we use the two most commonly reported ones: the OTIS/BETA/GAMMA and the California Test of Mental Maturity (CTMM). One problem with the NLS/1966 sample is that for each respondent we observe at most one of these test scores. For all respondents, there is a second cognitive test score. We use this test as a second measure of ability for all respondents in this sample, in addition to the measure that is available for each respondent.[10]

We model the test score j, M_j using specification (8). The covariates X^M include family background variables, year of birth dummies, and characteristics of the individuals at the time of the test.[11] To set the scale of θ_1, we normalize $\alpha_1^M = 1$.

One of the advantages of using factor models instead of the test score itself is that factor models allow for test scores to be noisy measures of cognitive skills. Another advantage of this method is that it does not require the observation of test scores for all individuals for its implementation. This is important because full samples exhibit different earnings characteristics than incomplete samples. Web Data Appendix Table 6 and Web Data Appendix Figures 1 through 4 compare the time series of the means and standard errors of earnings in the full NLSY/1979 sample and the NLSY/1979 subsample with observed test scores. While mean earnings are the same in both samples, the standard errors are more volatile in the subsample with observed test scores than in the full sample. Web Data Appendix Table 7 and Web Data Appendix Figures 5 through 9 make the same comparison for the NLS/1966, but the conclusions for this data set are different. Mean high-school earnings from age 35 to 41 tend to be higher in the subsample with observed test scores than in the full sample. The same is true for the time series for the standard error of

[10] See our web appendix for additional discussion of these tests.

[11] In both NLSY/1979 and NLS/1966 we include mother's education, father's education, number of siblings, urban residence at age 14, dummies for year of birth of the individuals, and an intercept. In the NLSY/1979 sample we also control for the fact that the test taker is enrolled in school and the highest grade completed at the time of the test. In the NLS/1966 all of the respondents were enrolled in school at the time of the test (in fact, the test score is obtained in a survey from schools). We do not know the highest grade completed at the time of the test for the NLS/1966 sample.

college earnings. Although there are no differences in mean college earnings, the standard errors diverge in the distinct samples, and they are much higher in the full sample than in the subsample with observed test scores (see Web Data Appendix Figure 8). Web Data Appendix Tables 8-10 compare the serial correlation matrices for NLSY/1979 in high-school, college and overall sample, respectively. Parallel information for NLS/1966 survey is reported in Web Data Appendix Tables 11-13. Although there are few differences in the serial correlation patterns when one compares the full sample with the subsample with observed test scores, the information contained in the subsample with observed test scores alone would not suffice to compute all the cells in the correlation matrix.

For the NLSY/1979, a six factor model fits the data best:

$$Y_{s,t} = X\beta_{s,t}+\theta_1\alpha_{1,s,t}+\theta_2\alpha_{2,s,t}+\theta_3\alpha_{3,s,t}+\theta_4\alpha_{4,s,t}+\theta_5\alpha_{5,s,t}+\theta_6\alpha_{6,s,t}+\varepsilon_{s,t},\ t=1,\ldots,T^*,\ s=0,1, \quad (18)$$

where $t = 1$ corresponds to age 22 and T^* is age 41. For the NLS/1966, only a five factor model is required to fit the data.[12] The identification of the model requires the normalization of some of the factor loadings. Web Supplement Appendix Table 1A shows the factor loading normalizations imposed in the NLSY/1979.[13] Web Supplement Appendix Table 1B shows the same information for the NLS/1966. In both samples, the covariates X are urban residence at age 14, dummies for year of birth of the individual, and an intercept.

The cost function C for the 1979 sample is

$$C = Z\gamma + \theta_1\alpha_{1,C} + \theta_2\alpha_{2,C} + \theta_3\alpha_{3,C} + \theta_4\alpha_{4,C} + \theta_5\alpha_{5,C} + \theta_6\alpha_{6,C} + \varepsilon_C. \quad (19)$$

The covariates Z are urban residence at age 14; dummies for year of birth; an intercept; and variables that affect the costs of going to college but do not affect outcomes $Y_{s,t}$ after controlling for ability, such as mother's education, father's education, number of siblings, and local tuition. Because we only have earnings data into the early 40's for both samples, the truncated discounted earnings after the 40's are absorbed into the definition of C. There is one fewer factor in the model fit on the 1966 sample.

[12] In the next subsection and at our website, we discuss the goodness-of-fit measures used to select the appropriate model for each sample.

[13] The Web Supplement Appendix is Part II of the Web Appendix and is distinct from the Data Description Appendix, Part II.

Each factor θ_k is assumed to be generated by a mixture of J_k normal distributions,

$$\theta_k \sim \sum_{j=1}^{J_k} p_{k,j} \phi\left(\theta_k \mid \mu_{k,j}, \lambda_{k,j}\right),$$

where $\phi\left(\eta \mid \mu_j, \lambda_j\right)$ is a normal density for η with mean μ_j and variance λ_j and $\sum_{j=1}^{J_k} p_{k,j} = 1$, and $p_{k,j} > 0$. Ferguson (1983) shows that mixtures of normals with a large number of components approximate any distribution of θ_k arbitrarily well in the ℓ^1 norm. The $\varepsilon_{s,t}$ are also assumed to be generated by mixtures of normals. We estimate the model using Markov Chain Monte Carlo methods as described in Carneiro, Hansen, and Heckman (2003). For all factors, a three-component model ($J_k = 3$, $k = 1, \ldots, 6$) is adequate. For all $\varepsilon_{s,t}$ we use a four-component model.[14]

3.1 How the model fits the data

The model fits the data well. Figure 1 compares actual and predicted densities of earnings for the overall sample for the NLSY/1979. The fit is good overall and in detailed subsamples. See Web Supplement Appendix Figures 1.1-3.40.[15] When we perform formal tests of equality of predicted versus actual densities, we pass these tests for most of the ages (see Web Supplement Appendix Table 2A for the NLSY/1979 and Web Supplement Appendix Table 2B for the NLS/1966). The model fits the NLS/1966 data marginally better than it fits the NLSY/1979 data.

We also perform χ^2 goodness-of-fit tests for the earnings correlation matrices. Table 1 shows that the six factor model fits the correlation matrix for the NLSY/1979 sample. We cannot reject the equality of actual and predicted correlation matrix for the NLS/1966 model when we use a five factor model. However, a five factor model does not fit the earnings correlation matrix for the NLSY/1979. Consequently, in what follows, we use a six factor model for the NLSY/1979 and a five factor model for NLS/1966.[16]

[14] Additional components do not improve the goodness of fit of the model to the data.

[15] The Web Supplement Appendix shows fits for all ages, for the overall, high-school, and college earnings, for both the NLSY/1979 and NLS/1966.

[16] Figures 4.1-4.11 in the Web Supplement Appendix plot the estimated densities of the factors for the NLS 1966 and 1979 NLSY samples by attained schooling level. Unsurprisingly, there is selection by education on the first three factors in both samples and no selection on the higher numbered factors.

3.2 The Evolution of Joint Distributions and Returns to College

In estimating the distribution of earnings in counterfactual schooling states within a policy regime (e.g., the distributions of college earnings for people who actually choose to be high school graduates under a particular tuition policy), one standard approach is to assume that college and high school distributions are the same except for an additive constant—the coefficient of a schooling dummy in an earnings regression possibly conditioned on the covariates. We relax this assumption and identify the joint distribution of counterfactuals without imposing this condition or other strong assumptions used in the literature.[17]

We identify both *ex ante* and *ex post* joint distributions. Let $E\left(Y_s|\mathcal{I}\right)$ denote the *ex ante* present value of lifetime earnings at schooling level s. Suppose that we want to compute the means and the covariances between *ex ante* college and *ex ante* high-school earnings conditional on information set $\mathcal{I}$, which we estimate. For a three factor case, the *ex ante* mean present value of earnings is

$$E\left(Y_s|\mathcal{I}\right)=\sum_{t=1}^{T^*}\frac{X\beta_{s,t}+\theta_1\alpha_{1,s,t}+\theta_2\alpha_{2,s,t}+\theta_3\alpha_{3,s,t}}{(1+\rho)^{t-1}},$$

where T^* is the maximum age at which we observe earnings. To simplify notation, the first age we analyze (age 22) is denoted $t=1$ and the last age we analyze (age 41) is denoted T^*. Conditional on covariates X, the covariance between $E\left(Y_1|\mathcal{I}\right)$ and $E\left(Y_0|\mathcal{I}\right)$ is

$$\begin{aligned}\mathrm{Cov}\left(E\left(Y_1|\mathcal{I}\right),E\left(Y_0|\mathcal{I}\right)\right) &= \mathrm{Var}\left(\theta_1\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{1,1,t}}{(1+\rho)^{t-1}}\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{1,0,t}}{(1+\rho)^{t-1}}\right)\\ &\quad+\cdots+\mathrm{Var}\left(\theta_3\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{3,1,t}}{(1+\rho)^{t-1}}\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{3,0,t}}{(1+\rho)^{t-1}}\right).\end{aligned}$$

Tables 2A and 2B present the conditional distributions of the present values of *ex ante* college earnings given *ex ante* high school earnings decile by decile for the NLSY/1979 and NLS/1966 samples, respectively. If the dependence across outcomes were perfect and positive, the diagonal elements would be '1' and the off diagonal elements would be '0.' We estimate positive dependence between the relative positions of individuals in the two distributions, but the dependence is far from perfect. For example, for the NLSY/1979 sample, 29.95% of the individuals who are in the first

[17] Abbring and Heckman (2007) discuss a variety of alternative assumptions used to identify joint counterfactual distributions.

decile of the high school present value of earnings distribution would be in the first decile of the college present value of earnings distribution. For the NLS/1966 sample, this figure is 70.36%. The comparison of tables 2A and 2B shows that the correlation between *ex ante* high school and *ex ante* college present value of lifetime earnings weakens in the more recent cohort.

We can also compute the covariance between the present value of *ex post* college and high-school earnings conditional on X. For the NLSY/1979 sample, this is

$$\begin{aligned}\operatorname{Cov}(Y_1, Y_0 | X) &= \operatorname{Var}(\theta_1)\left(\sum_{t=1}^{T^*}\frac{\alpha_{1,1,t}}{(1+\rho)^{t-1}}\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{1,0,t}}{(1+\rho)^{t-1}}\right) \\ &\quad + \cdots + \operatorname{Var}(\theta_6)\left(\sum_{t=1}^{T^*}\frac{\alpha_{6,1,t}}{(1+\rho)^{t-1}}\right)\left(\sum_{t=1}^{T^*}\frac{\alpha_{6,0,t}}{(1+\rho)^{t-1}}\right).\end{aligned}$$

Tables 3A and 3B show the conditional distributions of the present values of *ex post* college earnings given *ex post* high school earnings for the NLSY/1979 and NLS/1966 samples, respectively.[18] In NLSY/1979, *ex post* present values of earnings exhibit greater correlation than do present values of *ex ante* earnings (the correlation is 0.16 for *ex ante* earnings and 0.28 for *ex post* earnings). On the other hand, in the NLS/1966 sample, *ex post* earnings exhibit lower correlation than *ex ante* earnings (the correlation is 0.91 for ex-ante and 0.62 for *ex post* earnings.)

Knowledge of the joint distributions allows us to compare factual with counterfactual distributions. Take agents who choose to be high-school graduates. We can compare the density of the present value of *ex post* earnings in the high-school sector with those in the college sector for the people who are high-school graduates. This information is plotted in Figures 2A and 2B for the NLSY/1979 and NLS/1966, respectively. For both data sets, the high-school agents would have higher earnings if they had chosen to be college graduates. For college graduates, we compare the actual density of present value of earnings in the college sector with that in the high-school sector. We display these densities in Figures 3A and 3B for the NLSY/1979 and NLS/1966, respectively. Again, in both data sets the densities of high-school present value of earnings is to the left of the college density.

From such distributions we can generate the distribution of rates of return to college, where we define the *ex post* gross rate of return R (excluding cost) as $R = \frac{Y_1 - Y_0}{Y_0}$. The typical high

[18] Recall that the model for the NLS/1966 sample only requires five factors, so the last term in the preceding expression is deleted for that model.

school student would have returns of around 29% for a college education over the whole life cycle for the NLS/1966 sample and around 31% for the NLSY/1979 sample. For the typical college graduate, this return is around 33% for the NLS/1966 sample and 40% for the NLSY/1979 sample. For individuals at the margin, these figures are 31% and 35% for the NLS/1966 and NLSY/1979 samples, respectively. Returns to college have increased for college graduates and individuals at the margin, but not so much for the high school graduates.

From knowledge of the joint distribution, we can compute the percentage of individuals who regret their schooling choice. This is reported in Table 5. A higher fraction of the individuals who stop at high-school regret not graduating from college (7.5% in NLSY/1979 and 9.7% in NLS/1966). Around 3% of individuals who attend college regret not stopping their schooling upon high-school graduation, for both the NLSY/1979 and NLS/1966.

3.3 The Evolution of Uncertainty and Heterogeneity

The valuation or net utility function for schooling is

$$I = E\left(\sum_{t=1}^{T^*} \frac{Y_{1,t} - Y_{0,t}}{(1+\rho)^{t-1}} \middle| \mathcal{I}\right) - E\left(C \middle| \mathcal{I}\right).$$

Individuals go to college if $I > 0$. As explained in section 2.6, the correlation between schooling choices and future information allows us to disentangle heterogeneity from uncertainty. In the NLSY/1979, we test, and do not reject, the hypothesis that, at the time they make college going decisions, individuals know their Z and the factors θ_1, θ_2, and θ_3. They do not know the cohort dummies in X and the factors θ_4, θ_5, θ_6, or $\varepsilon_{s,t}$, $s = 0, 1$, $t = 1, \ldots, T^*$, at the time they make their educational choices. For the NLS/1966 we test, and do not reject, the hypothesis that the individuals know their Z, X, and the factors θ_1, θ_2, and θ_3. They do not know the cohort dummies in X and the factors θ_4, θ_5, or $\varepsilon_{s,t}$, $s = 0, 1$, $t = 1, \ldots, T^*$, at the time they make their educational choices. Thus, components not in the information sets of the agents at age 18 are different in the NLSY/1979 and in the NLS/1966. We now explore the implications of these estimates for the growth of uncertainty in the American economy.

3.3.1 Total Residual Variance and Variance of Unforecastable Component

For the model fit on NLSY/1979 data, the present value of lifetime (i.e., from age 22 ($t = 1$) to age 41 (T^*)) realized earnings in school level s can be written as

$$Y_s = \sum_{t=1}^{T^*} \frac{Y_{s,t}}{(1+\rho)^{t-1}} = \sum_{t=1}^{T^*} \frac{X\beta_{s,t} + \theta_1\alpha_{1,s,t} + \theta_2\alpha_{2,s,t} + \theta_3\alpha_{3,s,t} + \theta_4\alpha_{4,s,t} + \theta_5\alpha_{5,s,t} + \theta_6\alpha_{6,s,t} + \varepsilon_{s,t}}{(1+\rho)^{t-1}}.$$

We define the residual in the realized present value of earnings as the sum of the unobserved (by the econometrician) components,[19]

$$Q_s = \sum_{t=1}^{T^*} \frac{\theta_1\alpha_{1,s,t} + \theta_2\alpha_{2,s,t} + \theta_3\alpha_{3,s,t} + \theta_4\alpha_{4,s,t} + \theta_5\alpha_{5,s,t} + \theta_6\alpha_{6,s,t} + \varepsilon_{s,t}}{(1+\rho)^{t-1}}. \tag{20}$$

This term combines terms that are known and unknown by the agent at the time schooling choices are made. The total residual variance in schooling level s is $\text{Var}(Q_s)$.

The unforecastable component of the residual is the sum of the components that are not in the information set of the agent at the time schooling choices are made. For the NLSY/1979, the unforecastable component is

$$P_s = \sum_{t=1}^{T^*} \frac{\theta_4\alpha_{4,s,t} + \theta_5\alpha_{5,s,t} + \theta_6\alpha_{6,s,t} + \varepsilon_{s,t}}{(1+\rho)^{t-1}}. \tag{21}$$

The variance of the unforecastable component in schooling level s is $\text{Var}(P_s)$. Clearly $\text{Var}(P_s) \leq \text{Var}(Q_s)$.

Table 6A displays the total residual variance and the variance of the unforecastable components for each schooling level for both NLS/1966 (Panel A) and NLSY/1979 (Panel B). Total residual variance in present value of lifetime college earnings increase from 460.62 (NLS/1966) to 709.74 (NLSY/1979). This implies an increase of almost 55% in the total residual variance. The increase is larger for the present value of high school earnings: it goes from 284.80 in NLS/1966 to 507.29, corresponding to an increase of almost 80%.

The variance of the unforecastable component has also increased. For college earnings, it is 181.37 in the NLS/1966 and it becomes 372.35 in the NLSY/1979. For high school earnings, it is

[19] In our empirical analysis we fix $\rho = 0.05$.

128.43 in the NLS/1966 and becomes 272.35 in the NLSY/1979. In percentage terms, this implies that the variance of the unforecastable component increased 105% for college and 112% for high school.

We perform a similar analysis for the gross returns to college:

$$R = \sum_{t=1}^{T^*} \frac{Y_{1,t} - Y_{0,t}}{(1+\rho)^{t-1}}.$$

The total residual in the gross returns to college can be defined as $\Delta Q = Q_1 - Q_0$,

$$\Delta Q = \sum_{t=1}^{T^*} \frac{\theta_1 \Delta\alpha_{1,t} + \theta_2 \Delta\alpha_{2,t} + \theta_3 \Delta\alpha_{3,t} + \theta_4 \Delta\alpha_{4,t} + \theta_5 \Delta\alpha_{5,t} + \theta_6 \Delta\alpha_{6,t} + \Delta\varepsilon_t}{(1+\rho)^{t-1}},$$

and the unforecastable component in the gross returns to college is defined as $\Delta P = P_1 - P_0$,

$$\Delta P = \sum_{t=1}^{T^*} \frac{\theta_4 \Delta\alpha_{4,t} + \theta_5 \Delta\alpha_{5,t} + \theta_6 \Delta\alpha_{6,t} + \Delta\varepsilon_t}{(1+\rho)^{t-1}}.$$

Table 6A shows that total residual variance in gross returns to college increased from 351 in NLS/1966 to 906 in NLSY/1979, an increase of around 160%. The variance of the unforecastable components increased from 327 to 432, or roughly 32%.

This evidence shows that the increase in the variance of the unforecastable components of earnings is a key element in explaining the increase in the total residual variance in high school and college earnings. Furthermore, both the total residual variance and the variance of unforecastable components have increased more for low-skill workers (i.e., high-school graduates) than high-skill workers (i.e., college graduates). A similar exercise can be repeated to determine the evolution of unobserved heterogeneity for which, for both the NLSY/1979 and NLS/1966, the unobserved heterogeneity component is $\sum_{t=1}^{T^*} \frac{\theta_1 \alpha_{1,s,t} + \theta_2 \alpha_{2,s,t} + \theta_3 \alpha_{3,s,t}}{(1+\rho)^{t-1}}$.

Figures 5A and 5B plot the density of total residual versus the density of unforecastable components for high-school earnings for the 1979 and 1966 samples, respectively. Unforecastable components are more tightly dispersed in both samples. Figures 6A and 6B make the analogous comparison for college earnings for the 1979 and 1966 samples, respectively. Finally, Figures 7A and 7B show the corresponding figures for returns. Figure 7B reveals that in the 1966 sample there is very little

predictability in returns.

Table 6B presents the total residual variance and the variance of heterogeneity (or forecastable) components for each schooling level for both NLS/1966 (Panel A) and NLSY/1979 (Panel B). In the recent cohort, individuals have become more diverse in predictable ways. For college earnings, the variance of forecastable components is 279 for the NLS/1966. It is 337 for the NLSY/1979, which corresponds to a roughly 21% increase. For high school earnings, it is 156 for the NLS/1966 and 234 for the NLSY/1979, which implies an increase of more than 50%. As is evident from Figure 4B, there is little selection on returns in the NLS/1966. This happens because agents could not forecast returns well in the 1966 cohort and most of the variance of unobservable component in returns for that cohort is due to uncertainty and not forecastable heterogeneity. (See Figure 7B.) There is a substantial increase in the variance of heterogeneity in the returns to college for the more recent cohort. In summary, this analysis shows that about 75% of the increase in the variability in college wages, 65% of the increase in the variability in high school wages, and about 20% of the increase in the variability of returns to college is due to an increase in uncertainty in the American labor market. We next turn to an analysis of how the increase in variance is apportioned by age.

3.3.2 The Variance of the Unforecastable Component by Age

The increase in uncertainty is not uniform across age. For every age t and schooling level s let $P_{s,t}$ denote the unforecastable component in school level s age t earnings. Our estimates, along with the identifying normalizations displayed in Table 1 of the Web Supplement Appendix, imply that the unforecastable components for ages 22 through 25 for the 1979 cohort are given by

$$P_{s,t} = \frac{\varepsilon_{s,t}}{(1-\rho)^{t-1}} \text{ for } t = 1, \ldots, 4, \tag{22}$$

and for ages 26 through 41 by

$$P_{s,t} = \frac{\theta_4 \alpha_{4,s,t} + \theta_5 \alpha_{5,s,t} + \theta_6 \alpha_{6,s,t} + \varepsilon_{s,t}}{(1+\rho)^{t-1}} \text{ for } t = 5, \ldots, T^*. \tag{23}$$

Figure 8 plots the variance of unforecastable components in high school earnings in NLS/1966 and NLSY/1979. They are about the same until age 27/28. From age 29 on, the variances diverge. They

both increase with age, but the NLSY/1979 cohort experiences a more rapid increase in variances with age than does the NLS/1966 cohort. At age 41, the variance of the unforecastable component in high school earnings for the NLSY/1979 cohort is almost three times larger than its counterpart in the NLS/1966 sample.

A similar pattern appears in the variances of the unforecastable components in college earnings. Figure 9 shows that until around age 30, the profiles of the variances are roughly the same for the NLSY/1979 and NLS/1966 cohorts. From age 31 on, the series diverge, and the variances in the NLSY/1979 sample increase at a faster rate. At age 37, the variances of the unforecastable component in NLSY/1979 are more than twice those in NLS/1966 sample.

3.3.3 Accounting for Macro Uncertainty

Our estimates of uncertainty are microeconomic in nature. The literature in macroeconomics documents that aggregate instability has decreased in the past 30 years (see Gordon, 2005). To capture this phenomenon, we introduce time dummies into the earnings equation. Given the standard problem of the lack of identification of age, period, and cohort effects, we cannot identify cohort effects in the presence of age and time effects.[20] We find that the variables that capture macro uncertainty (time dummies for earnings equations after schooling choices are made) do not enter the schooling choice equation. Thus, we estimate that macro uncertainty is not forecastable by agents at the time schooling choices are made. However, realized macro shocks affect earnings outcomes. Macro uncertainty decreased for later cohorts by 90% (see Table 7). These estimates are consistent with the evidence that US business cycle volatility has decreased in recent years. At the same time, macro uncertainty is a tiny fraction of total uncertainty for both cohorts (5% for 1966; 1% for 1979).

4 Sequential Revelation of Information, More General Preferences and Market Settings

To focus on identifying agent information sets, we analyze a one-shot model of schooling choices. We also assume risk neutrality. This allows us to use expected present value income maximization

[20] See Heckman and Robb (1985) for a discussion of this problem and a discussion of the interactions that can be identified.

as our schooling choice criterion. A basic question is "What can be identified in more general environments?" In the absence of perfect certainty or perfect risk sharing, preferences and market environments also determine schooling choices. The separation theorem used in this paper that allows consumption and schooling decisions to be analyzed in isolation of each other breaks down.

If we postulate information arrival processes *a priori*, and assume that preferences are known up to some unknown parameters as in Flavin (1981), Blundell and Preston (1998) and Blundell, Pistaferri, and Preston (2004), we can identify departures from specified market structures.[21,22] An open question, not yet fully resolved in the literature, is how far one can go in nonparametrically jointly identifying preferences, market structures and agent information sets.[23] One can add consumption data to the schooling choice and earnings data to secure identification of risk preference parameters (within a parametric family) and information sets, and to test among alternative models for market environments.[24] Alternative assumptions about what analysts know produce different interpretations of the same evidence. The lack of full insurance interpretation given to the empirical analysis by Flavin (1981) and Blundell, Pistaferri, and Preston (2004), may instead be a consequence of their misspecification of the generating processes of agent information sets.

5 Summary and Conclusion

This paper investigates the sources of rising wage inequality the US labor market. We find that increasing inequality arises from both increasing micro uncertainty and increasing heterogeneity predictable by agents. The latter could arise from increased sorting. Both components have increased since the late 1960s. The fraction of the variability due to micro uncertainty has increased. Aggregate uncertainty has decreased. Thus the recent increase of uncertainty has microeconomic origins. Our evidence of substantially increased uncertainty at the micro level for recent cohorts of unskilled labor supports the turbulence hypothesis of Ljungqvist and Sargent (2004).

[21] Flavin (1981), Blundell and Preston (1998) and Blundell, Pistaferri, and Preston (2002) specify explicit time series processes for the unobservables (e.g., ARMA or fixed effect/AR-1 models) with unknown coefficients but prespecified serial correlation structures and assume that the innovations in these processes are the uncertainty components while the predictable components are heterogeneity.

[22] Hansen (1987) shows a fundamental nonidentification result for the Flavin model estimated on aggregate data. Our use of micropanel data circumvents the problem he raises.

[23] This point was first made at the Hicks Lecture at Oxford, April 2004, and is published in Cunha, Heckman, and Navarro (2005).

[24] Navarro (2005) attempts to do this.

A Identification of the Model

We provide an intuitive discussion of identification based on normal errors. Normality joined with the assumption of expected value income maximization produces closed form solutions. See Carneiro, Hansen, and Heckman (2003) for proofs of semi-parametric identification of the distributions of the factors θ and uniquenesses ε without the normality assumption.

A.1 Test Scores

First consider identification of the test score equations. Test scores are available for all agents and are determined before they make their college decisions. There is no selection bias in the test score equations. Three assumptions are crucial in securing identification through factor models. First, the explanatory variables X^M are independent of θ_1 and ε_k^M, for $k = 1, \ldots, K$. Second, the factor θ_1 is independent of ε_k^M, for $k = 1, \ldots, K$. Third, the uniqueness ε_k^M is independent from ε_l^M for any $k \neq l$, for $k, l = 1, \ldots, K$. The first assumption, along with standard rank conditions, allows β_k^M to be consistently estimated from a simple OLS regression of M_k against X^M. Given the β_k^M, we can construct differences $M_k - X^M \beta_k^M$ and compute the covariances:

$$\text{Cov}\left(M_1 - X^M\beta_1^M, M_2 - X^M\beta_2^M\right) = \alpha_1^M \alpha_2^M \sigma_{\theta_1}^2, \tag{24}$$

$$\text{Cov}\left(M_1 - X^M\beta_1^M, M_3 - X^M\beta_3^M\right) = \alpha_1^M \alpha_3^M \sigma_{\theta_1}^2, \tag{25}$$

$$\text{Cov}\left(M_2 - X^M\beta_2^M, M_3 - X^M\beta_3^M\right) = \alpha_2^M \alpha_3^M \sigma_{\theta_1}^2. \tag{26}$$

The left-hand sides of (24), (25), and (26) can be computed from the data. The right-hand sides of (24), (25), and (26) are implied by the factor model. As is common in the factor literature, we need to normalize one of the factor loadings to set the scale of the factor. Let $\alpha_1^M = 1$. If we take the ratio of (26) to (24) we identify α_3^M. Analogously, the ratio of (26) to (25) allows us to recover α_2^M. Given the normalization of $\alpha_1^M = 1$ and identification of α_2^M, we identify $\sigma_{\theta_1}^2$ from (24). Finally, we can identify the variance of ε_k^M from the variance of $M_k - X^M\beta_k^M$. Because the factor θ_1 and uniquenesses ε_k are independently normally distributed random variables, we have identified their distribution.

A.2 Earnings and Choice Equations

To establish identification of the objects of interest in earnings equations requires a little more work because of the selection problem. Our assumption of normally distributed factors and uniquenesses simplifies the analysis because we can use closed-form solutions to reduce the identification problem to the identification of a few parameters.

We rely on four key assumptions to secure identification. First, all of the observable explanatory variables X and Z are independent of the unobservable factors, θ_1 and θ_2, as well as uniquenesses $\varepsilon_{s,t}$ for all s, t. Second, θ_1 is independent of θ_2. Third, both θ_1 and θ_2 are independent of ε_C and $\varepsilon_{s,t}$ for all s, t. Fourth, $\varepsilon_{s,t}$ is independent from ε_C and $\varepsilon_{s',t'}$ for any pairs s, s' and t, t' such that $s \neq s'$ or $t \neq t'$. All of the dependence among $U_{0,t}, U_{1,t}$, and U_C is captured through the factors θ_1 and θ_2. As a consequence of these assumptions,

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2_{\theta_1} & 0 \\ 0 & \sigma^2_{\theta_2} \end{bmatrix} \right).$$

Because the loadings $\alpha_{1,s,t}$, $\alpha_{2,s,t}$, $\alpha_{1,C}$, and $\alpha_{2,C}$ can be freely specified, the factors θ can affect $U_{0,t}, U_{1,t}$, and U_C differently. The joint distribution of the labor earnings $Y_{0,t}, Y_{1,t}$ conditional on X is

$$\begin{bmatrix} Y_{0,t} \\ Y_{1,t} \end{bmatrix} \mid X \sim N\left(\begin{bmatrix} X\beta_{0,t} \\ X\beta_{1,t} \end{bmatrix}, \begin{bmatrix} \alpha^2_{1,0,t}\sigma^2_{\theta_1} + \alpha^2_{2,0,t}\sigma^2_{\theta_2} + \sigma^2_{\varepsilon_{0,t}} & \alpha_{1,0,t}\alpha_{1,1,t}\sigma^2_{\theta_1} + \alpha_{2,0,t}\alpha_{2,1,t}\sigma^2_{\theta_2} \\ \alpha_{1,0,t}\alpha_{1,1,t}\sigma^2_{\theta_1} + \alpha_{2,0,t}\alpha_{2,1,t}\sigma^2_{\theta_2} & \alpha^2_{1,1,t}\sigma^2_{\theta_1} + \alpha^2_{2,1,t}\sigma^2_{\theta_2} + \sigma^2_{\varepsilon_{1,t}} \end{bmatrix} \right). \tag{27}$$

As a result, identification of the joint distribution $F(Y_{0,t}, Y_{1,t} \mid X)$ reduces to the identification of the parameters $\beta_{s,t}, \alpha_{k,s,t}, \sigma_{\varepsilon_{s,t}}$, and $\sigma^2_{\theta_j}$ for $s = 0, 1$; $t = 1, \ldots, T$ and $j = 1, 2$, and $k = 1, 2$. From the observed data and the factor structure assumption it follows that

$$E(Y_{1,t} | X, S = 1) = X\beta_{1,t} + \alpha_{1,1,t}E[\theta_1 | X, S = 1] + \alpha_{2,1,t}E[\theta_2 | X, S = 1] + E[\varepsilon_{1,t} | X, S = 1]. \tag{28}$$

The event $S = 1$ corresponds to the event $I = E\left(\sum_{t=1}^{T} \left(\frac{1}{1+\rho} \right)^{t-1} (Y_{1,t} - Y_{0,t}) - C \middle| \mathcal{I} \right) \geq 0$. As-

suming that $\varepsilon_{s,t}$ does not enter agent information sets, for the case $\{\theta_1, \theta_2\} \subset \mathcal{I}$ we obtain

$$E\left(\sum_{t=1}^{T}\left(\frac{1}{1+\rho}\right)^{t-1}(Y_{1,t}-Y_{0,t})-C\middle|\mathcal{I}\right)=\mu_I(X,Z)+\alpha_{1,I}\theta_1+\alpha_{2,I}\theta_2-\varepsilon_C.$$

Let η be the linear combination of three independent normal random variables: $\eta = \alpha_{1,I}\theta_1 + \alpha_{2,I}\theta_2 - \varepsilon_C$. Then, $\eta \sim N\left(0, \sigma_\eta^2\right)$, with $\sigma_\eta^2 = \alpha_{1,I}^2\sigma_{\theta_1}^2 + \alpha_{2,I}^2\sigma_{\theta_2}^2 + \sigma_{\varepsilon_c}^2$ and

$$S = 1 \Leftrightarrow \eta > -\mu_I(X,Z). \tag{29}$$

If we replace (29) in (28) and use the fact that $\varepsilon_{s,t}$ is independent of X, Z, and θ,

$$E\left(Y_{1,t}\middle|\, X, S=1\right) = X\beta_1 + \alpha_{1,1,t}E\left[\theta_1\middle|\, X, \eta > -\mu_I(X,Z)\right] + \alpha_{2,1,t}E\left[\theta_2\middle|\, X, \eta > -\mu_I(X,Z)\right]. \tag{30}$$

Because θ_1, θ_2 and η are normal random variables,

$$\theta_j = \frac{\mathrm{Cov}\left(\theta_j, \eta\right)}{\mathrm{Var}\left(\eta\right)}\eta + \rho_j \text{ for } j = 1, 2, \tag{31}$$

where ρ_j is a mean zero, normal random variable independent from η. Because $\mathrm{Cov}\left(\theta_1, \eta\right) = \sigma_{\theta_1}^2\alpha_{1,I}$ and $\mathrm{Cov}\left(\theta_2, \eta\right) = \sigma_{\theta_2}^2\alpha_{2,I}$ it follows that

$$E\left[\theta_1\middle|\, X, \eta > -\mu_I(X,Z)\right] = \frac{\sigma_{\theta_1}^2\alpha_{1,I}}{\sigma_\eta^2}E\left[\eta\middle|\, \eta > -\mu_I(X,Z)\right]$$

and

$$E\left[\theta_2\middle|\, X, \eta > -\mu_I(X,Z)\right] = \frac{\sigma_{\theta_2}^2\alpha_{2,I}}{\sigma_\eta^2}E\left[\eta\middle|\, \eta > -\mu_I(X,Z)\right].$$

For any standard normal random variable μ, $E\left(\mu\middle|\, \mu \geq -c\right) = \frac{\phi(c)}{\Phi(c)}$ where $\phi\left(.\right)$ and $\Phi\left(.\right)$ are the density and distribution function of a standard normal random variable. Define, for $j = 0, 1$, $\pi_{j,t} = \left(\frac{\alpha_{1,j,t}\alpha_{1,I}\sigma_{\theta_1}^2 + \alpha_{2,j,t}\alpha_{2,I}\sigma_{\theta_2}^2}{\sigma_\eta}\right)$. These facts together allow us to rewrite (28) as

$$E\left(Y_{1,t}\middle|\, \eta \leq -\mu_I(X,Z)\right) = X\beta_{1,t} + \pi_{1,t}\frac{\phi\left(\frac{\mu_I(X,Z)}{\sigma_\eta}\right)}{\Phi\left(\frac{\mu_I(X,Z)}{\sigma_\eta}\right)}. \tag{32}$$

We can derive a similar expression for mean observed earnings in sector "0":

$$E\left(Y_{0,t}|\,\eta > -\mu_I(X,Z)\right) = X\beta_{0,t} - \pi_{0,t}\frac{\phi\left(\frac{\mu_I(X,Z)}{\sigma_\eta}\right)}{\Phi\left(\frac{\mu_I(X,Z)}{\sigma_\eta}\right)}. \tag{33}$$

We can apply the two-step procedure developed in Heckman (1976) to identify $\beta_{0,t}, \beta_{1,t}$, $\pi_{0,t}$ and $\pi_{1,t}$. Given identification of $\beta_{s,t}$ for all s and t, we can construct the differences $Y_{s,t} - X\beta_{s,t}$ and compute the covariances

$$\text{Cov}\left(M_1 - X^M\beta_1^M, Y_{0,t} - X\beta_{0,t}\right) = \alpha_{1,0,t}\sigma_{\theta_1}^2 \tag{34}$$

and

$$\text{Cov}\left(M_1 - X^M\beta_1^M, Y_{1,t} - X\beta_{1,t}\right) = \alpha_{1,1,t}\sigma_{\theta_1}^2. \tag{35}$$

The left-hand sides of (34) and (35) are identified from the data. The right-hand sides are an implication of the factor model. We determined $\sigma_{\theta_1}^2$ from the analysis of the test scores. From equations (34) and (35), we can recover $\alpha_{1,0,t}$ and $\alpha_{1,1,t}$ for all t. Note that we can also identify $\alpha_{1,C}/\sigma_\eta$ by computing the covariance

$$\text{Cov}\left(M_1 - X\beta_1^M, \frac{I - \mu_I(X,Z)}{\sigma_\eta}\right) = \frac{\sum_{t=1}^{T}\left(\frac{1}{1+\rho}\right)^{t-1}(\alpha_{1,1,t} - \alpha_{1,0,t}) - \alpha_{1,C}}{\sigma_\eta}\sigma_{\theta_1}^2. \tag{36}$$

Using (34) and (35), we can identify $\alpha_{1,1,t}$ and $\alpha_{1,0,t}$ for all t. The only remaining term to be identified is the ratio $\alpha_{1,C}/\sigma_\eta$, which can be identified from covariance equation (36).

Note that if $T \geq 2$, we can also identify the parameters related to factor θ_2, such as $\alpha_{2,s,t}$ and $\sigma_{\theta_2}^2$. To see this, first normalize $\alpha_{2,0,1} = 1$ and compute the covariances:

$$\text{Cov}\left(Y_{0,1} - X\beta_{0,1}, Y_{0,2} - X\beta_{0,2}\right) - \alpha_{1,0,1}\alpha_{1,0,2}\sigma_{\theta_1}^2 = \alpha_{2,0,2}\sigma_{\theta_2}^2, \tag{37}$$

$$\text{Cov}\left(Y_{0,1} - X\beta_{0,1}, \frac{I - \mu_I(X,Z)}{\sigma_\eta}\right) - \frac{\alpha_{1,0,1}\sigma_{\theta_1}^2\sum_{t=1}^{T}(\alpha_{1,1,t} - \alpha_{1,0,t} - \alpha_{1,C})}{\sigma_\eta} = \frac{\sigma_{\theta_2}^2\sum_{t=1}^{T}(\alpha_{2,1,t} - \alpha_{2,0,t} - \alpha_{2,C})}{\sigma_\eta}, \tag{38}$$

$$\text{Cov}\left(Y_{0,2}-X\beta_{0,2},\frac{I-\mu_I(X,Z)}{\sigma_\eta}\right)-\frac{\alpha_{1,0,2}\sigma^2_{\theta_1}\sum_{t=1}^{T}\left(\alpha_{1,1,t}-\alpha_{1,0,t}-\alpha_{1,C}\right)}{\sigma_\eta}=\frac{\alpha_{2,0,2}\sigma^2_{\theta_2}\sum_{t=1}^{T}\left(\alpha_{2,1,t}-\alpha_{2,0,t}-\alpha_{2,C}\right)}{\sigma_\eta}. \tag{39}$$

The left-hand sides of (37), (38), and (39) are identified from the data. Computing the ratio of (39) to (38), we can recover $\alpha_{2,0,2}$. From (37) we can recover $\sigma^2_{\theta_2}$. We now add in the information on the covariances from the college earnings equation:

$$\text{Cov}\left(Y_{1,1}-X\beta_{1,1},Y_{1,2}-X\beta_{1,2}\right)-\alpha_{1,1,1}\alpha_{1,1,2}\sigma^2_{\theta_1}=\alpha_{2,1,1}\alpha_{2,1,2}\sigma^2_{\theta_2}, \tag{40}$$

$$\text{Cov}\left(Y_{1,1}-X\beta_{1,1},\frac{I-\mu_I(X,Z)}{\sigma_\eta}\right)-\frac{\alpha_{1,1}\sigma^2_{\theta_1}\sum_{t=1}^{T}\left(\alpha_{1,1,t}-\alpha_{1,0,t}-\alpha_{1,C}\right)}{\sigma_\eta}=\frac{\alpha_{2,1,1}\sigma^2_{\theta_2}\sum_{t=1}^{T}\left(\alpha_{2,1,t}-\alpha_{2,0,t}-\alpha_{2,C}\right)}{\sigma_\eta}, \tag{41}$$

$$\text{Cov}\left(Y_{1,2}-X\beta_{1,2},\frac{I-\mu_I(X,Z)}{\sigma_\eta}\right)-\frac{\alpha_{1,1,2}\sigma^2_{\theta_1}\sum_{t=1}^{T}\left(\alpha_{1,1,t}-\alpha_{1,0,t}-\alpha_{1,C}\right)}{\sigma_\eta}=\frac{\alpha_{2,1,2}\sigma^2_{\theta_2}\sum_{t=1}^{T}\left(\alpha_{2,1,t}-\alpha_{2,0,t}-\alpha_{2,C}\right)}{\sigma_\eta}. \tag{42}$$

Computing the ratios of (42) to (40) and (41) to (40), we obtain $\alpha_{2,1,2}$ and $\alpha_{2,1,1}$ respectively. Finally, we use the information available from the data on earnings by schooling choice, $\text{Var}\left(Y_{0,t}|\,X,S=0\right)$ and $\text{Var}\left(Y_{1,t}|\,X,S=1\right)$, to compute $\sigma^2_{\varepsilon_{0,t}}$ and $\sigma^2_{\varepsilon_{1,t}}$, respectively. Note that we have identified all of the elements that characterize the joint distribution as specified in (27).

The identification analysis in this Appendix uses the data on test scores in an essential way. However, with sufficiently long panel earnings data, it is possible to identify the model without test score data. See the analysis in Abbring and Heckman (2007).

References

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics, Volume 6*. Amsterdam: Elsevier. Forthcoming.

Blundell, R., L. Pistaferri, and I. Preston (2002). Partial insurance, information and consumption dynamics. Technical Report IFS Working Papers: W02/16, Institute for Fiscal Studies, London.

Blundell, R., L. Pistaferri, and I. Preston (2004, October). Consumption inequality and partial insurance. Technical Report WP04/28, Institute for Fiscal Studies.

Blundell, R. and I. Preston (1998, May). Consumption inequality and income uncertainty. *Quarterly Journal of Economics 113*(2), 603–640.

Carneiro, P., K. Hansen, and J. J. Heckman (2003, May). Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college choice. *International Economic Review 44*(2), 361–422. 2001 Lawrence R. Klein Lecture.

Carroll, C. D. (1994, February). How does future income affect current consumption? *Quarterly Journal of Economics 109*(1), 111–147.

Cunha, F., J. J. Heckman, and S. Navarro (2005, April). Separating uncertainty from heterogeneity in life cycle earnings, the 2004 Hicks lecture. *Oxford Economic Papers 57*(2), 191–261.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In H. Chernoff, M. Rizvi, J. Rustagi, and D. Siegmund (Eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, pp. 287–302. New York: Academic Press.

Flavin, M. A. (1981, October). The adjustment of consumption to changing expectations about future income. *Journal of Political Economy 89*(5), 974–1009.

Gordon, R. J. (2005). What caused the decline in U. S. business cycle volatility? In C. Kent and D. Norman (Eds.), *The Changing Nature of the Business Cycle*, pp. 61–104. Sydney, Australia:

Economics Group, Reserve Bank of Australia. Proceedings of a conference held at the H.C. Coombs Centre for Financial Studies, Kirribilli, Australia on 11-12 July 2005.

Gottschalk, P. and R. Moffitt (1994). The growth of earnings instability in the U.S. labor market. *Brookings Papers on Economic Activity 2*, 217–254.

Hansen, L. P. (1987). Calculating asset prices in three example economies. In T. F. Bewley (Ed.), *Advances in Econometrics: Fifth World Congress*, Volume 1, pp. 207–243. New York: Cambridge University Press.

Heckman, J. J. (1976, December). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement 5*(4), 475–492.

Heckman, J. J. and R. Robb (1985). Using longitudinal data to estimate age, period and cohort effects in earnings equations. In W. M. Mason and S. E. Fienberg (Eds.), *Cohort Analysis in Social Research: Beyond the Identification Problem*. New York: Springer-Verlag.

Heckman, J. J. and J. Scheinkman (1987, April). The importance of bundling in a Gorman-Lancaster model of earnings. *Review of Economic Studies 54*(2), 243–355.

Hill, M. S., G. J. Duncan, and P. V. Marsden (1992). *The Panel Study of Income Dynamics: A User's Guide*. Newbury Park, CA: Sage Publications.

Katz, L. F. and D. H. Autor (1999). Changes in the wage structure and earnings inequality. In O. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Volume 3, Chapter 25, pp. 1463–1555. New York: North-Holland.

Levy, F. and R. J. Murnane (1992, September). U.S. earnings levels and earnings inequality: A review of recent trends and proposed explanations. *Journal of Economic Literature 30*(3), 1333–1381.

Ljungqvist, L. and T. J. Sargent (2004, April-May). European unemployment and turbulence revisited in a matching model. *Journal of the European Economic Association 2*(2-3), 456–468.

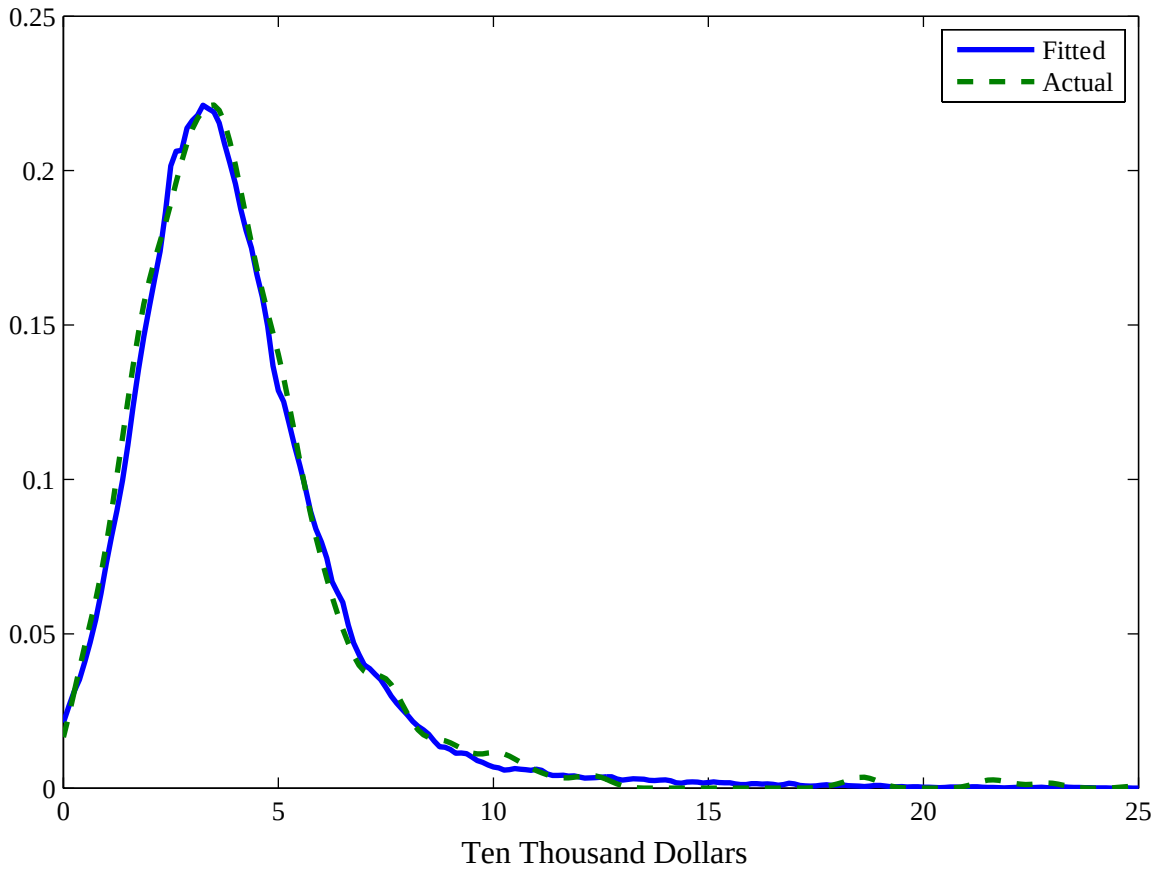
Matzkin, R. L. (1992, March). Nonparametric and distribution-free estimation of the binary threshold crossing and the binary choice models. *Econometrica 60*(2), 239–270.

Miller, S. (2004). *The National Longitudinal Surveys NLSY79 User's Guide 1979-2002*. Washington, DC: Bureau of Labor Statistics, U.S. Department of Labor.

Navarro, S. (2005). *Understanding Schooling: Using Observed Choices to Infer Agent's Information in a Dynamic Model of Schooling Choice When Consumption Allocation is Subject to Borrowing Constraints*. Ph.D. Dissertation, University of Chicago, Chicago, IL.

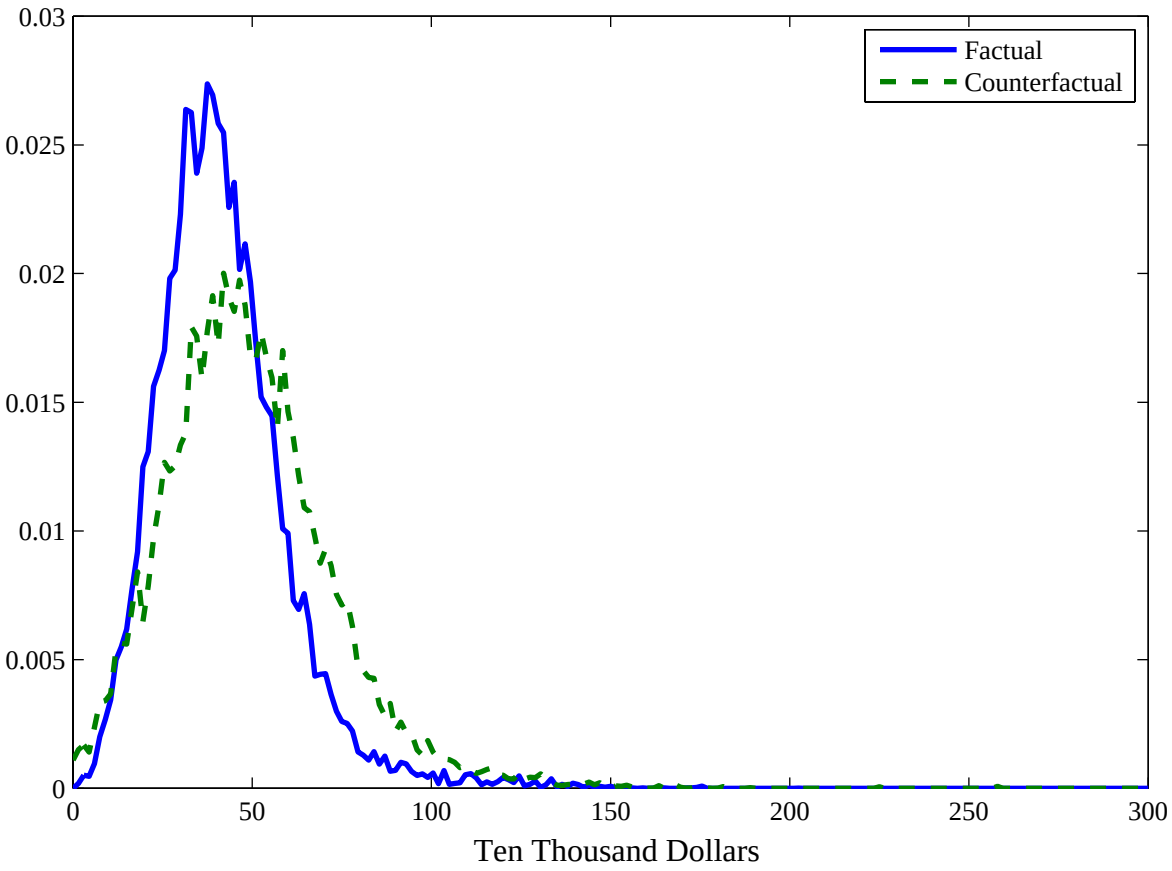
Sims, C. A. (1972, September). Money, income, and causality. *American Economic Review 62*(4), 540–552.

Figure 1
Densities of earnings at age 31
Overall Sample NLSY/1979



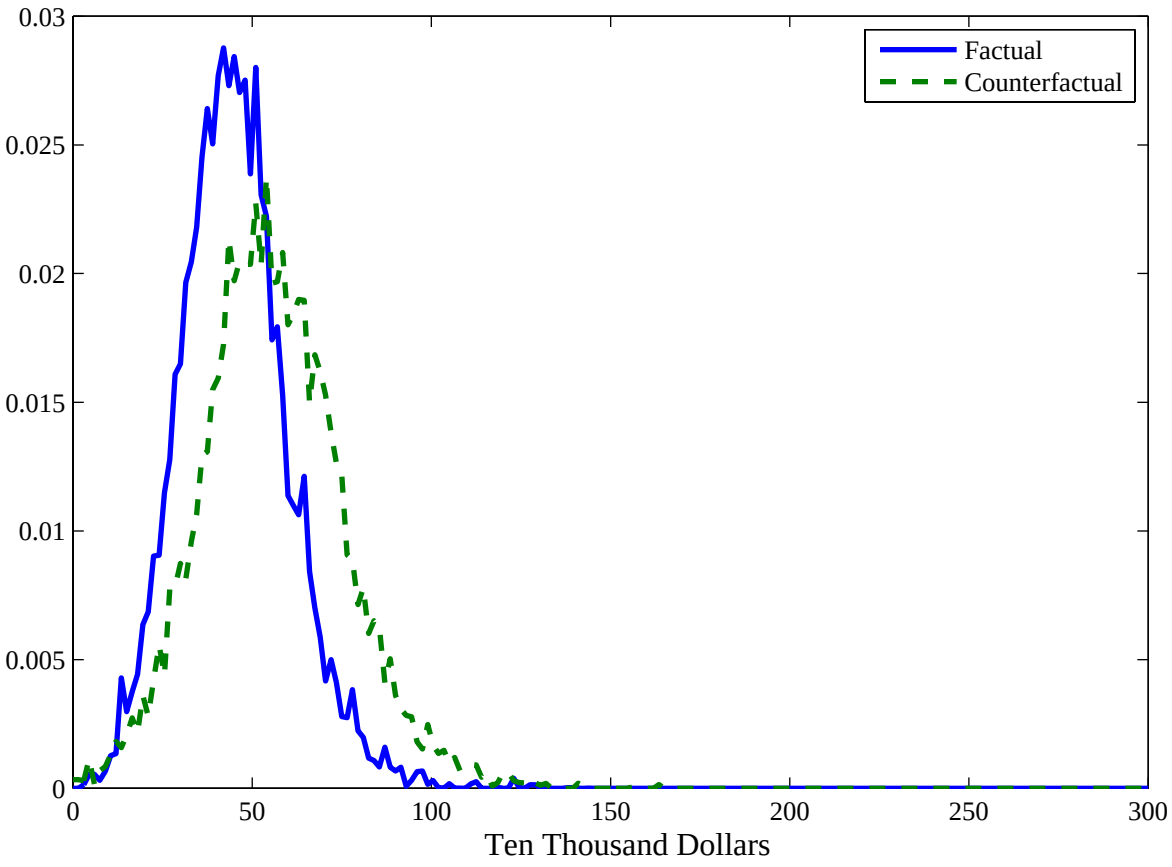
Let Y denote earnings at age 31 in the overall sample. Here we plot the density functions f(y) generated from the data (the solid curve), against that predicted by the model (the dashed line).

Figure 2A
Densities of present value of earnings
High School Sample NLSY/1979



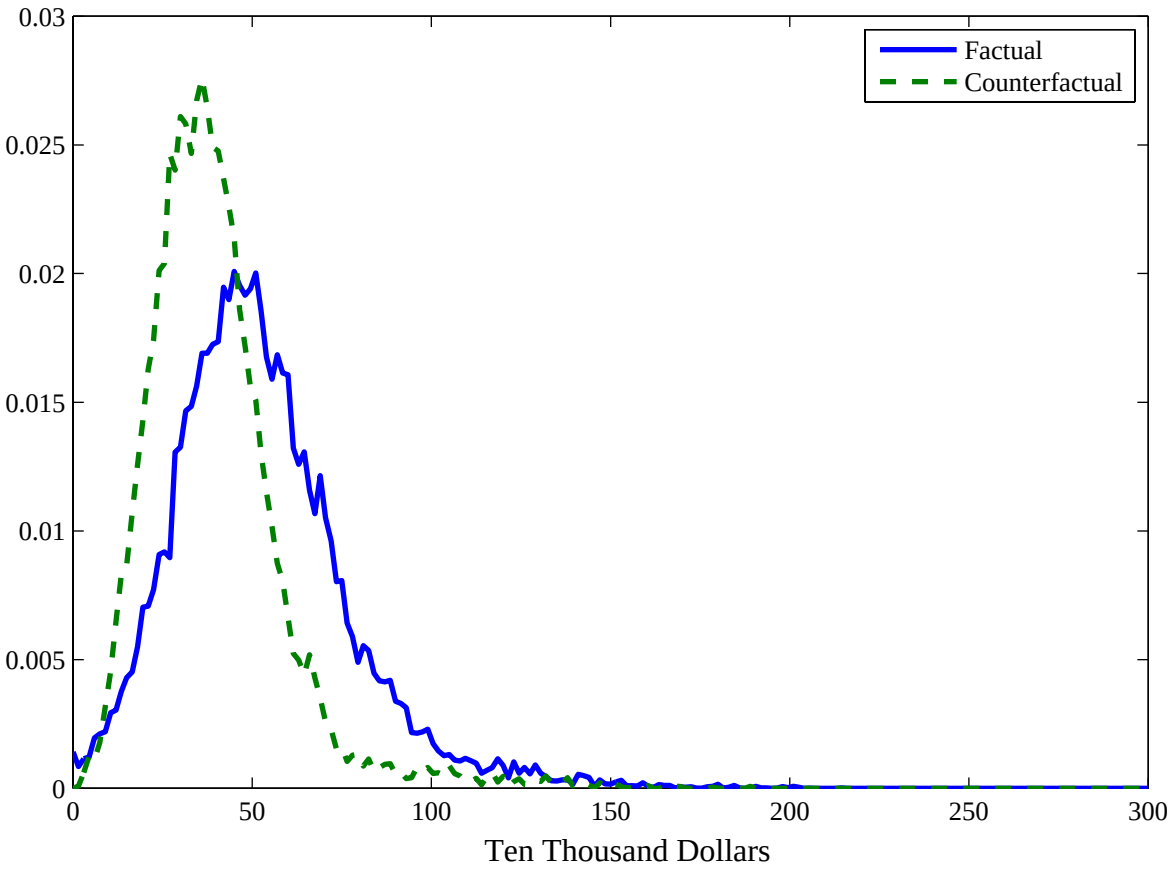
Let Y_0 denote the present value of earnings from age 22 to 41 in the High School sector (S = 0). Let Y_1 denote the present value of earnings from age 22 to 41 in the college sector (S = 1). Here we plot the factual density function $f(y_0|S=0)$ (the solid curve) against the counterfactual density function $f(y_1|S=0)$ (the dashed curve). We use a discount rate of 5%.

Figure 2B
Densities of present value of earnings
High School Sample NLS/1966



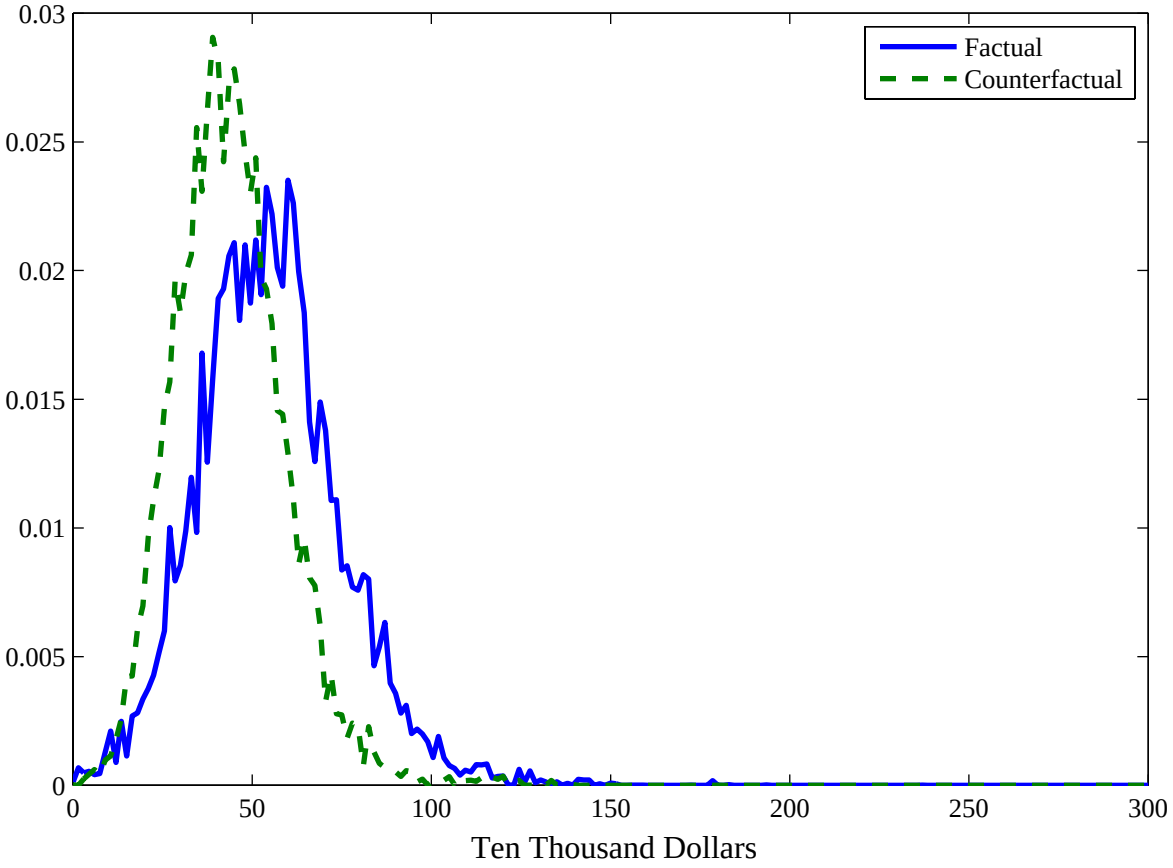
Let Y_0 denote the present value of earnings from age 22 to 41 in the High School sector (S = 0). Let Y_1 denote the present value of earnings from age 22 to 41 in the college sector (S = 1). Here we plot the factual density function $f(y_0|S=0)$ (the solid curve) against the counterfactual density function $f(y_1|S=0)$ (the dashed curve). We use a discount rate of 5%.

Figure 3A
Densities of present value of earnings
College Sample NLSY/1979



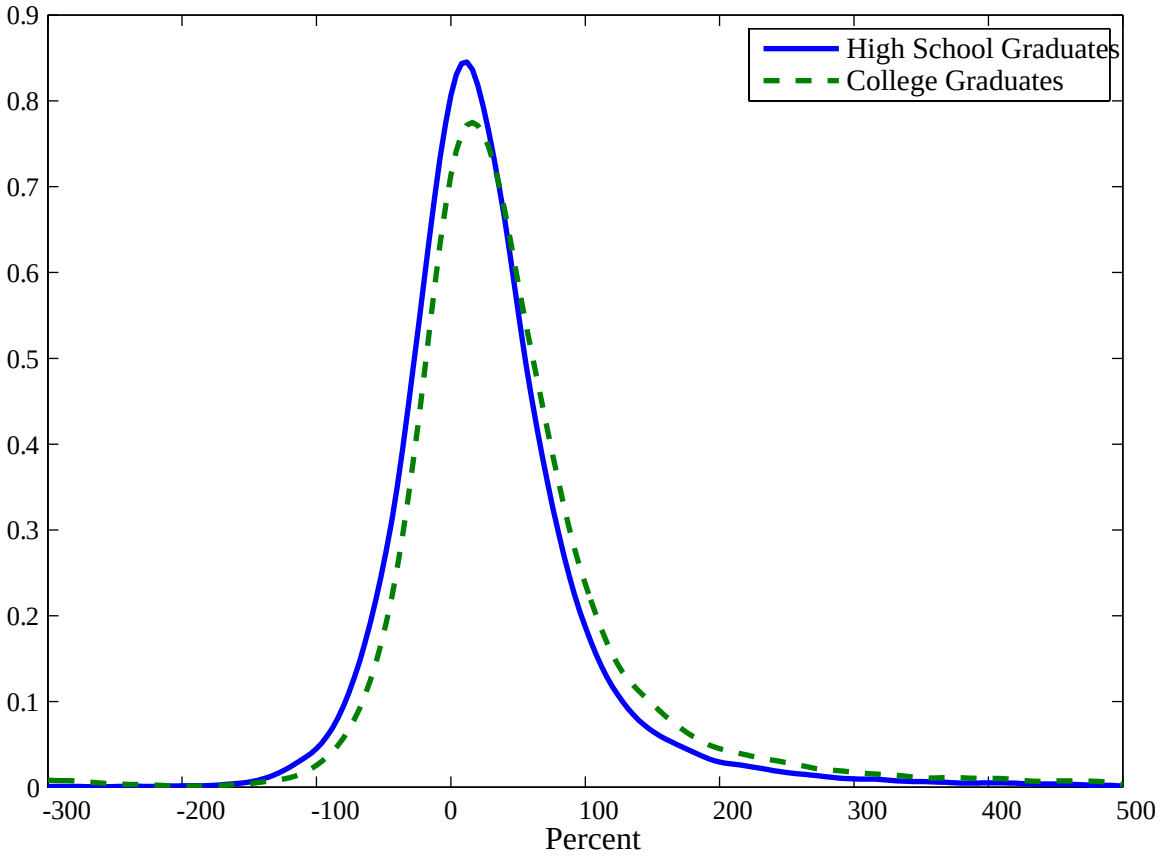
Let Y_0 denote the present value of earnings from age 22 to 41 in the High School sector (S = 0). Let Y_1 denote the present value of earnings from age 22 to 41 in the college sector (S = 1). Here we plot the factual density function $f(y_1|S=1)$ (the solid curve) against the counterfactual density function $f(y_0|S=1)$ (the dashed curve). We use a discount rate of 5%.

Figure 3B
Densities of present value of earnings
College Sample NLS/1966



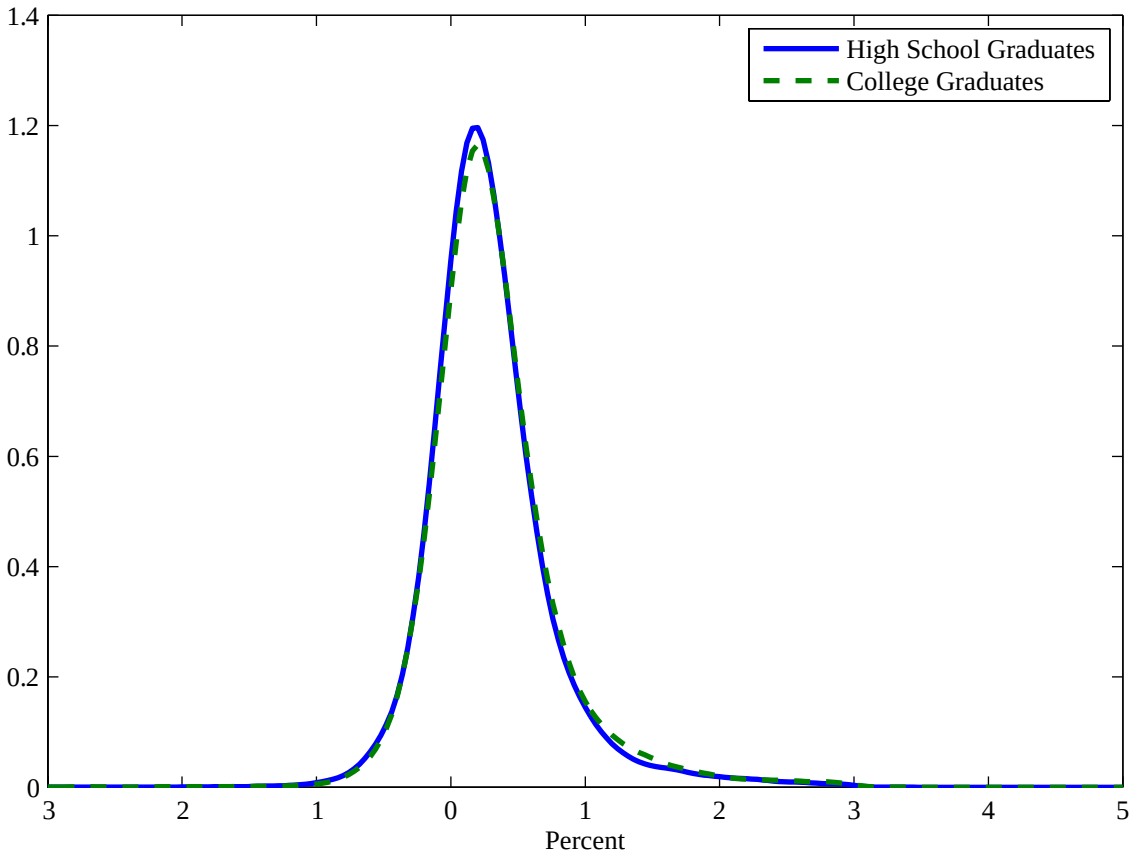
Let Y_0 denote the present value of earnings from age 22 to 41 in the High School sector (S = 0). Let Y_1 denote the present value of earnings from age 22 to 41 in the college sector (S = 1). Here we plot the factual density function $f(y_1|S=1)$ (the solid curve) against the counterfactual density function $f(y_0|S=1)$ (the dashed curve). We use a discount rate of 5%.

Figure 4A
Densities of Returns to College
NLSY/1979 Sample



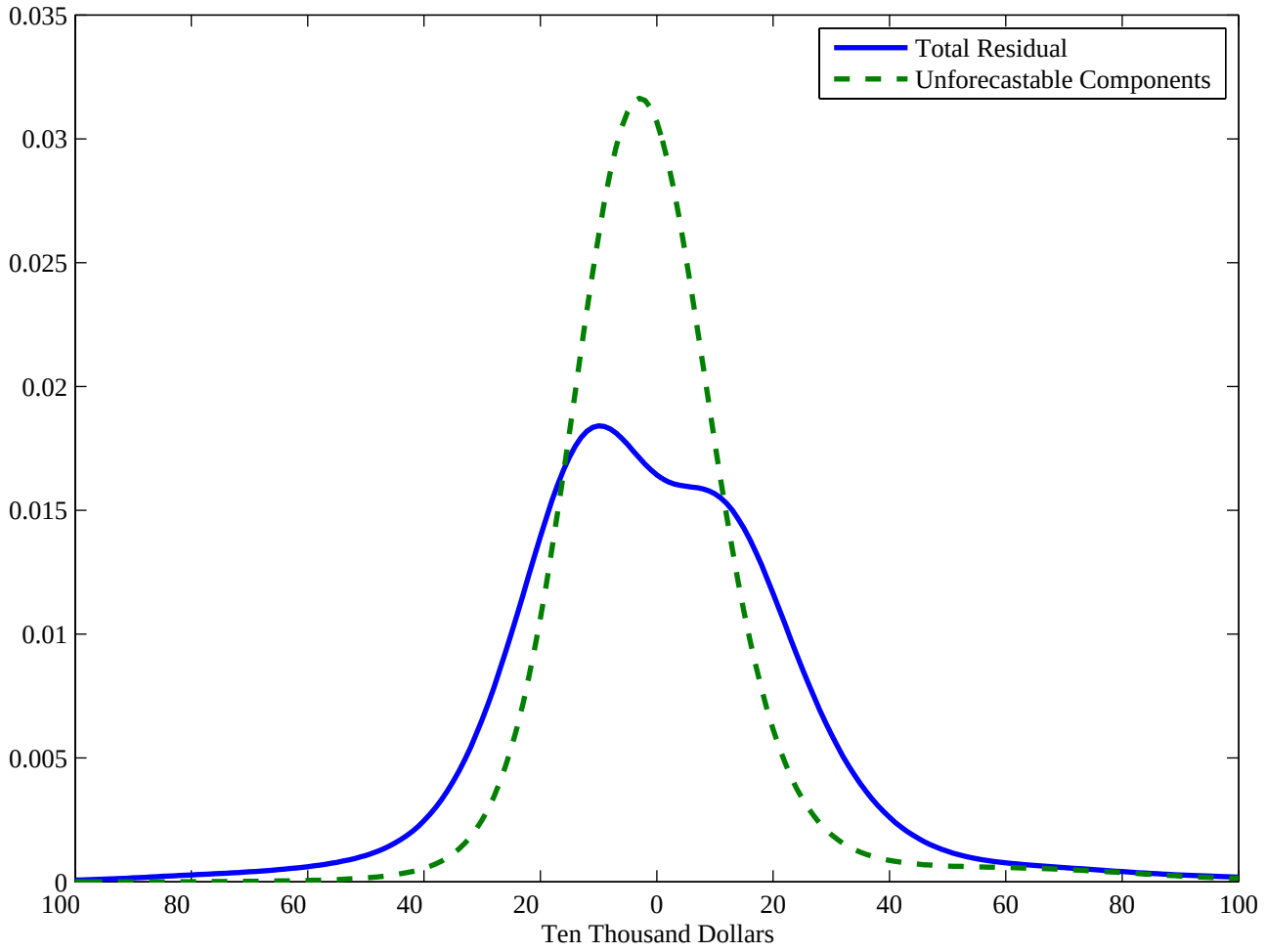
Let Y_0, Y_1 denote the present value of earnings from age 22 to age 41 in the high school and college sectors, respectively. Define ex post returns to college as the ratio $R=(Y_1-Y_0)/Y_0$. Let f(r) denote the density function of the ex post returns to college R. The solid line is the density of ex post returns to colege for high school graduates, that is, f(r|S=0). The dashed line is the density of ex post returns to college for college graduates, that is, f(r|S=1).

Figure 4B
Densities of Returns to College
NLS/1966 Sample



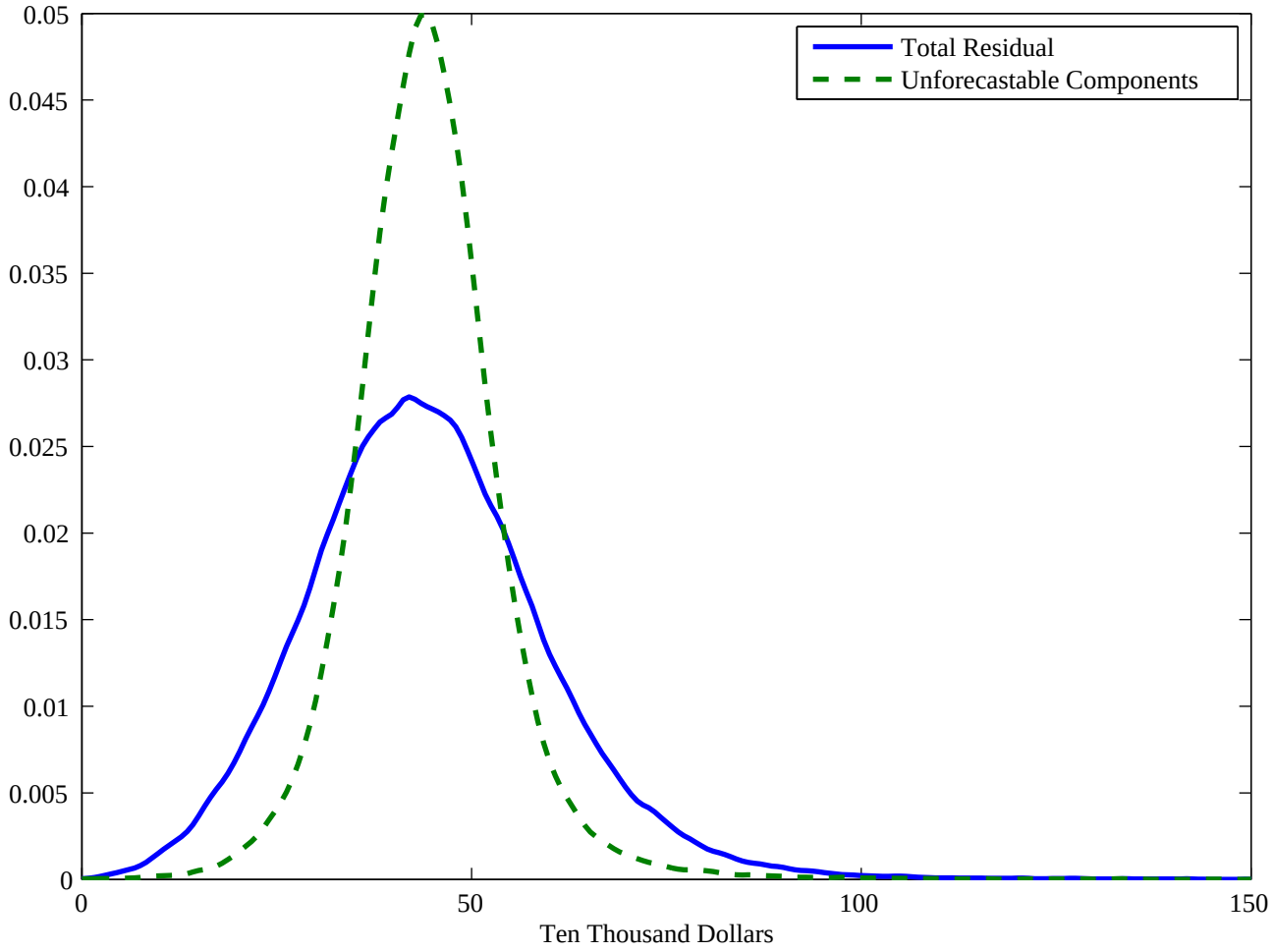
Let Y_0, Y_1 denote the present value of earnings from age 22 to age 41 in the high school and college sectors, respectively. Define ex post returns to college as the ratio $R=(Y_1-Y_0)/Y_0$. Let f(r) denote the density function of the ex post returns to college R. The solid line is the density of ex post returns to colege for high school graduates, that is, f(r|S=0). The dashed line is the density of ex post returns to college for college graduates, that is, f(r|S=1).

Figure 5A
The densities of total residual vs unforecastable components
in present value of high school earnings for the NLSY/1979 sample



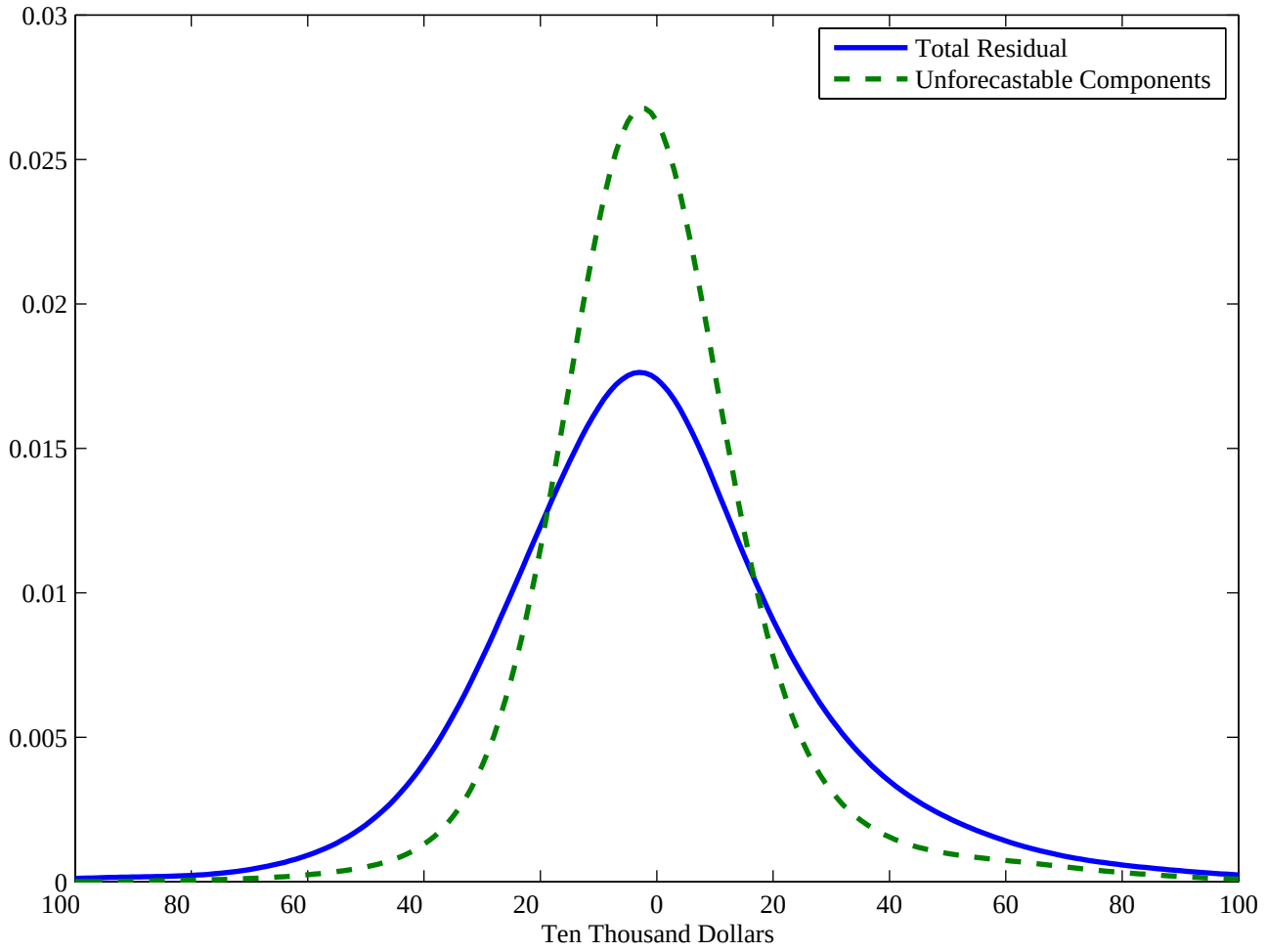
In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of high-school earnings from ages 22 to 41 for the NLSY/1979 sample of white males. The present value of earnings is calculated using a 5% interest rate.

Figure 5B
The densities of total residual vs unforecastable components
in present value of high school earnings for the NLS/1966 sample



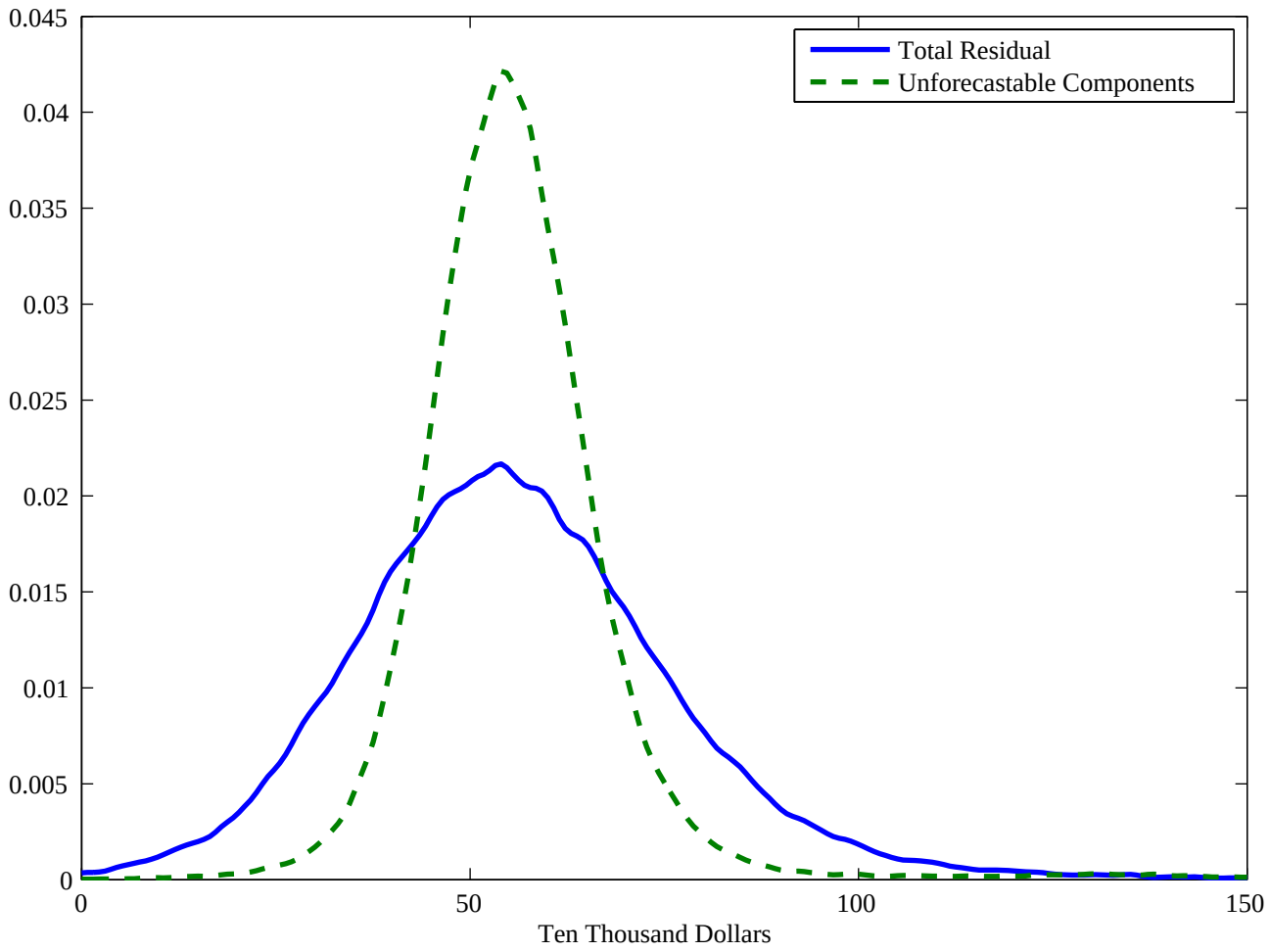
In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of high-school earnings from ages 22 to 41 for the NLS/1966 sample of white males. The present value of earnings is calculated using a 5% interest rate.

Figure 6A
The densities of total residual vs unforecastable components
in present value of college earnings for the NLSY/1979 sample



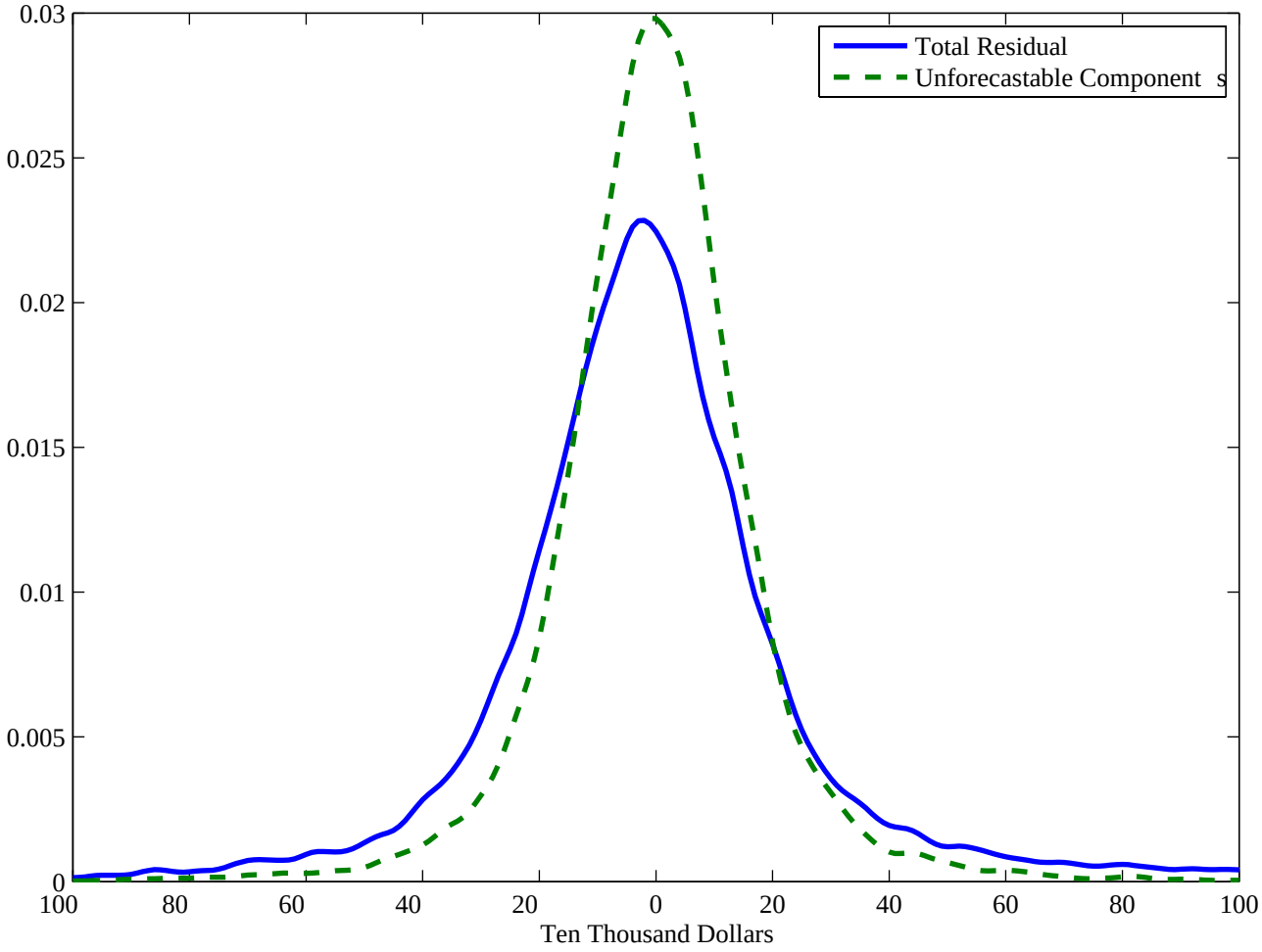
In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of college earnings from ages 22 to 41 for the NLSY/1979 sample of white males. The present value of earnings is calculated using a 5% interest rate.

Figure 6B
The densities of total residual vs unforecastable components
in present value of college earnings for the NLS/1966 sample



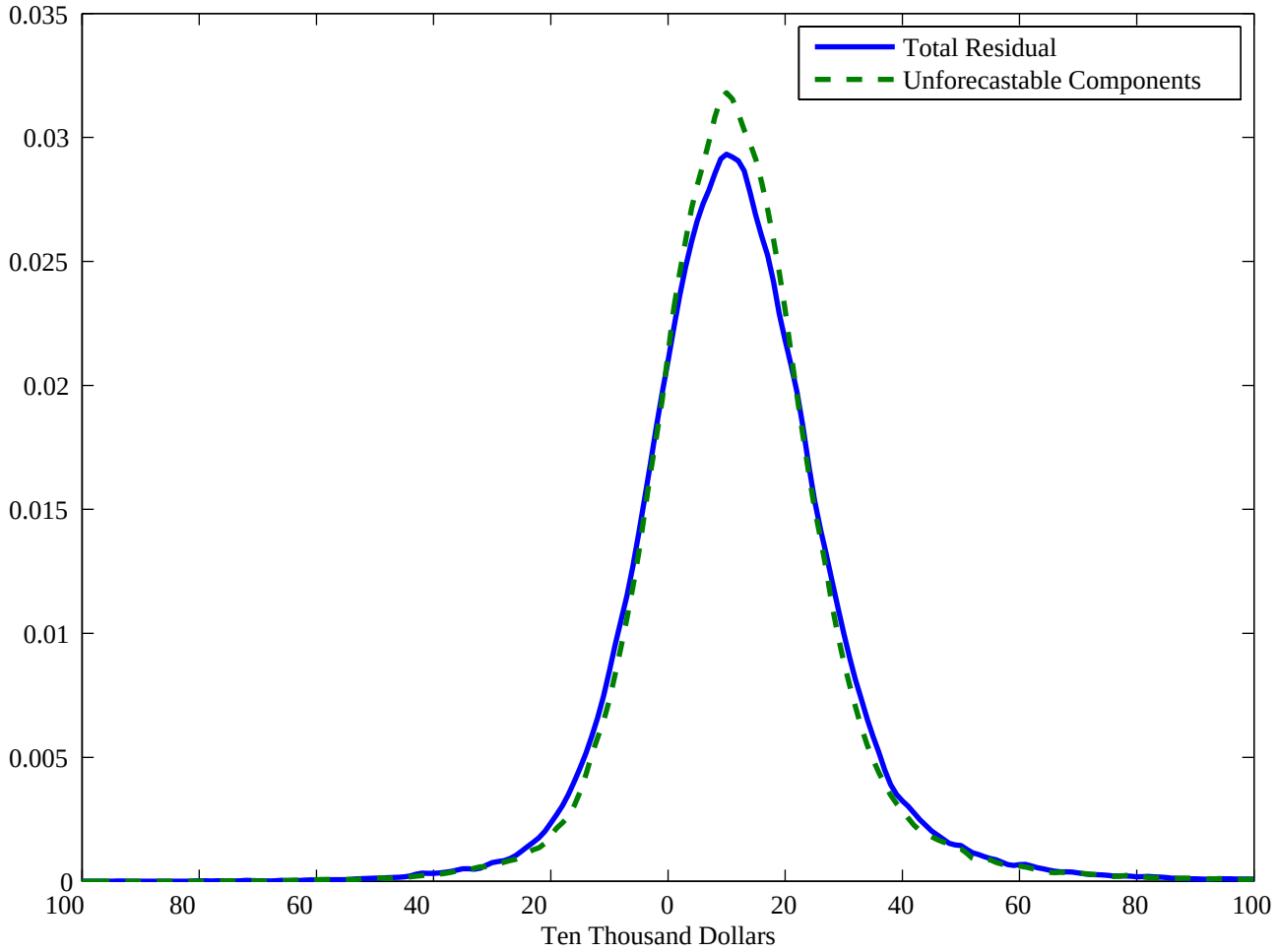
In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of college earnings from ages 22 to 41 for the NLS/1966 sample of white males. The present value of earnings is calculated using a 5% interest rate.

Figure 7A
Densities of total residual vs unforecastable components
returns to college vs high school for the NLSY/1979 sample



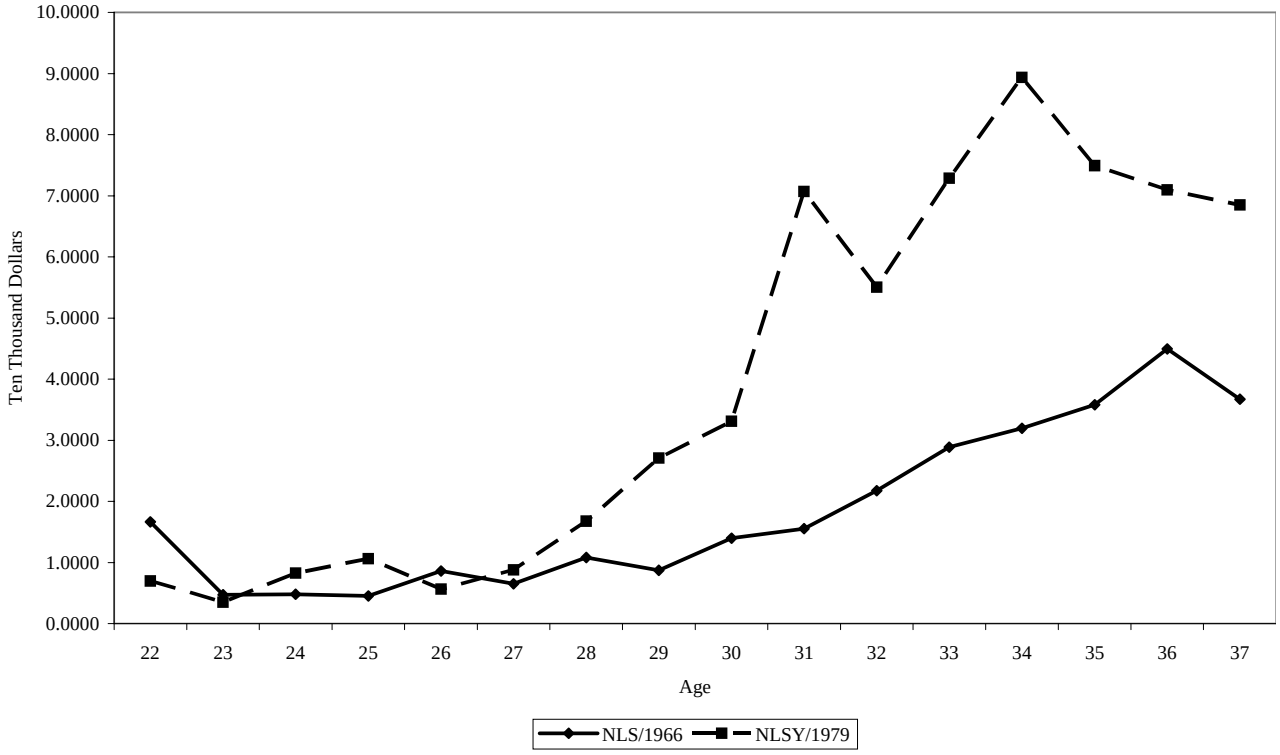
In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of earnings differences (or returns to college) for the white males sample of the NLSY/1979 from ages 22 to 41. The present value of returns to college is calculated using a 5% interest rate.

Figure 7B
Densities of total residual vs unforecastable components
returns to college vs high school for the NLS/1966 sample



In this figure we plot the density of total residual (the solid curve) against the density of the unforecastable components (the dashed curve) for the present value of earnings differences (or returns to college) for the white males sample of the NLSY/1979 from ages 22 to 41. The present value of returns to college is calculated using a 5% interest rate.

Figure 8
Evolution of Variance of Unforecastable Components - High School Sector

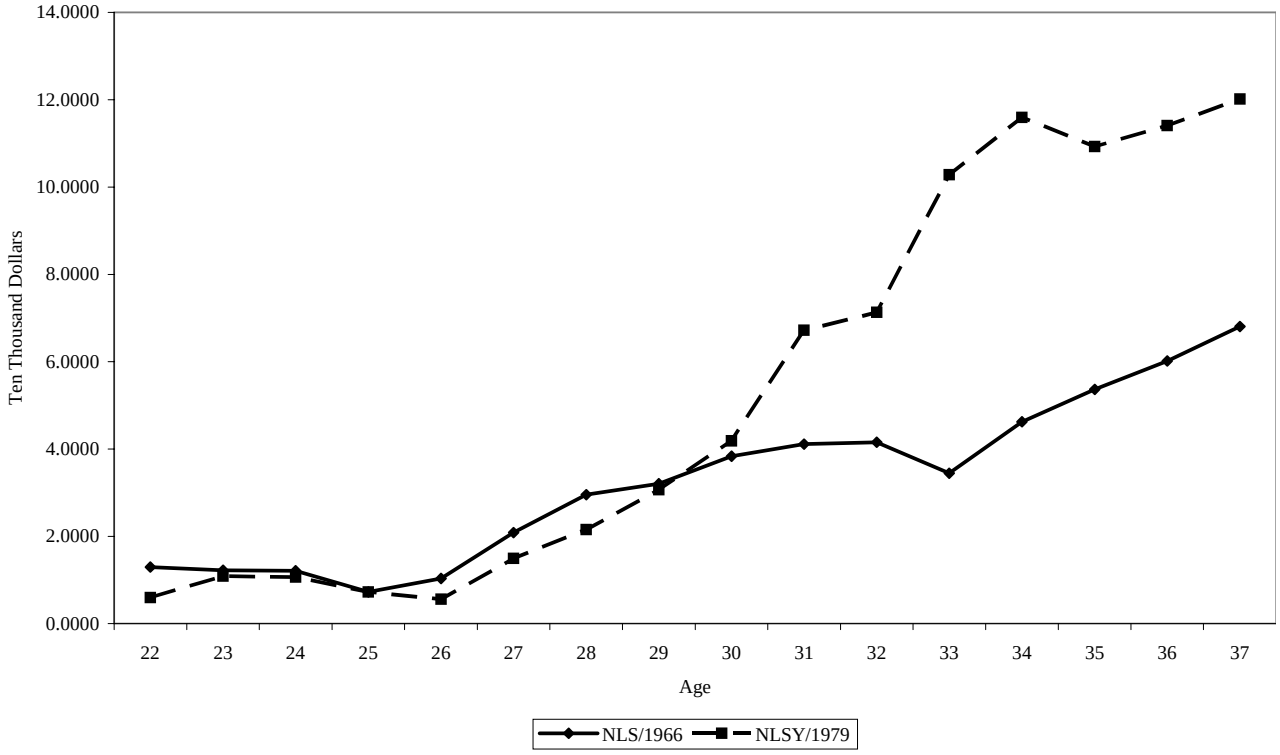


For each schooling level s, at each age t, we model earnings $Y_{s,t}$ according to:

$$Y_{s,t} = X\beta_{s,t} + \theta\alpha_{s,t} + \varepsilon_{s,t}$$

For the NLS/1966 data set, the vector θ contains 5 elements. We test and cannot reject that the agents know the factors θ_1, θ_2, and θ_3 but they don't know factors θ_4, θ_5, and $\varepsilon_{s,t}$ at the time of their schooling choice, for $s = 0, 1$ and $t = 22, ..., 41$. For the NLSY/1979 data set, the vector θ contains 6 elements. We test and cannot reject that the NLSY/1979 respondents know the factors θ_1, θ_2, and θ_3 but they don't know factors $\theta_4, \theta_5, \theta_6$ and $\varepsilon_{s,t}$ at the time of their schooling choice, for $s = 0, 1$ and $t = 22, ..., 41$. Let $P_{s,t}$ denote the unforecastable components at the time of the schooling choice. For the NLS/1966, $P_{s,t} = \alpha_{4,s,t}\theta_4 + \alpha_{5,s,t}\theta_5 + \varepsilon_{s,t}$. For the NLSY/1979, $P_{s,t} = \alpha_{4,s,t}\theta_4 + \alpha_{5,s,t}\theta_5 + \alpha_{6,s,t}\theta_6 + \varepsilon_{s,t}$. In Figure 10, we compare the variance of $P_{s,t}$ from NLS/1966 (the solid curve) with the one from NLSY/1979 (the dashed curve) at different ages of the individuals who are high-school graduates. We see that until age 27, the estimated variance of $P_{s,t}$ from NLS/1966 and NLSY/1979 are very similar, but from age 28 on, the variance of $P_{s,t}$ from NLSY/1979 is much larger than the counterpart from NLS/1966.

Figure 9
Evolution of Variance of Unforecastable Components - College Sector



For each schooling level s, at each age t, we model earnings $Y_{s,t}$ according to:

$$Y_{s,t} = X\beta_{s,t} + \theta\alpha_{s,t} + \varepsilon_{s,t}$$

For the NLS/1966 data set, the vector θ contains 5 elements. We test and cannot reject that the agents know the factors θ_1, θ_2, and θ_3 but they don't know factors θ_4, θ_5, and $\varepsilon_{s,t}$ at the time of their schooling choice, for $s = 0, 1$ and $t = 22, ..., 41$. For the NLSY/1979 data set, the vector θ contains 6 elements. We test and cannot reject that the NLSY/1979 respondents know the factors θ_1, θ_2, and θ_3 but they don't know factors $\theta_4, \theta_5, \theta_6$ and $\varepsilon_{s,t}$ at the time of their schooling choice, for $s = 0, 1$ and $t = 22, ..., 41$. Let $P_{s,t}$ denote the unforecastable components at the time of the schooling choice. For the NLS/1966, $P_{s,t} = \alpha_{4,s,t}\theta_4 + \alpha_{5,s,t}\theta_5 + \varepsilon_{s,t}$. For the NLSY/1979, $P_{s,t} = \alpha_{4,s,t}\theta_4 + \alpha_{5,s,t}\theta_5 + \alpha_{6,s,t}\theta_6 + \varepsilon_{s,t}$. In Figure 11, we compare the variance of $P_{s,t}$ from NLS/1966 (the solid curve) with the one from NLSY/1979 (the dashed curve) at different ages of the individuals who are college graduates. We see that until age 30, the estimated variance of $P_{s,t}$ from NLS/1966 and NLSY/1979 are very similar, but from age 31 on, the variance of $P_{s,t}$ from NLSY/1979 is much larger than the counterpart from NLS/1966.

Table 1

Test of Equality of Predicted versus Actual Correlation Matrices of Earnings (from ages 22 to 41) NLSY/1979 and NLS/1966

	High School	College	Overall
NLS/1966 - 5 Factors	15.6968	210.4133	114.8754
NLS/1979 - 6 Factors	70.6451	156.5446	187.5425
NLS/1979 - 5 Factors	64.2682	309.2815	226.2401
Critical Value*	222.0741	222.0741	222.0741

* 95% Confidence

Table 2A: Ex-Ante Conditional Distributions for the NLSY/1979 (College Earnings Conditional on High School Earnings)
$Pr(d_i<Yc<d_i+1 \mid d_j<Yh<d_j+1)$ where d_i is the ith decile of the College Lifetime Ex-Ante Earnings Distribution and d_j is the jth decile of the High School Ex-Ante Lifetime Earnings Distribution
Individual fixes unknown θ at their means, so Information Set=$\{\theta_1,\theta_2,\theta_3\}$
Corrrelation(Y_C,Y_H) = 0.1666

	College									
High School	1	2	3	4	5	6	7	8	9	10
1	0.2995	0.1685	0.1114	0.0789	0.0570	0.0413	0.0393	0.0431	0.0471	0.1137
2	0.2273	0.2119	0.1597	0.1271	0.0907	0.0678	0.0450	0.0288	0.0180	0.0236
3	0.1532	0.1840	0.1656	0.1472	0.1146	0.0914	0.0642	0.0434	0.0230	0.0132
4	0.1110	0.1368	0.1492	0.1474	0.1418	0.1184	0.0882	0.0588	0.0334	0.0148
5	0.0748	0.1100	0.1244	0.1413	0.1459	0.1403	0.1172	0.0836	0.0462	0.0162
6	0.0494	0.0866	0.1146	0.1204	0.1371	0.1399	0.1283	0.1242	0.0736	0.0258
7	0.0306	0.0582	0.0904	0.1094	0.1264	0.1436	0.1506	0.1430	0.1064	0.0414
8	0.0236	0.0348	0.0531	0.0769	0.0989	0.1252	0.1638	0.1799	0.1676	0.0761
9	0.0264	0.0262	0.0316	0.0459	0.0651	0.0929	0.1308	0.1784	0.2431	0.1594
10	0.0457	0.0182	0.0214	0.0216	0.0321	0.0446	0.0772	0.1176	0.2291	0.3925

Table 2B: Ex-Ante Conditional Distributions for the NLS/1966 (College Earnings Conditional on High School Earnings)
$Pr(d_i<Yc<d_i+1 \mid d_j<Yh<d_j+1)$ where d_i is the ith decile of the College Lifetime Ex-Ante Earnings Distribution and d_j is the jth decile of the High School Ex-Ante Lifetime Earnings Distribution
Individual fixes unknown θ at their means, so Information Set=$\{\theta_1,\theta_2,\theta_3\}$
Corrrelation(Y_C,Y_H) = 0.9174

	College									
High School	1	2	3	4	5	6	7	8	9	10
1	0.7036	0.2155	0.0622	0.0137	0.0035	0.0015	0.0000	0.0000	0.0000	0.0000
2	0.2225	0.3780	0.2475	0.1085	0.0285	0.0110	0.0035	0.0000	0.0005	0.0000
3	0.0500	0.2505	0.2960	0.2320	0.1090	0.0455	0.0120	0.0035	0.0015	0.0000
4	0.0145	0.1005	0.2250	0.2585	0.2150	0.1135	0.0545	0.0135	0.0045	0.0005
5	0.0045	0.0435	0.1055	0.1945	0.2545	0.2135	0.1265	0.0460	0.0105	0.0010
6	0.0010	0.0115	0.0435	0.1190	0.2035	0.2455	0.2100	0.1335	0.0295	0.0030
7	0.0000	0.0030	0.0150	0.0500	0.1190	0.2185	0.2705	0.2095	0.1040	0.0105
8	0.0005	0.0000	0.0055	0.0200	0.0555	0.1085	0.2080	0.3125	0.2460	0.0435
9	0.0000	0.0000	0.0005	0.0035	0.0105	0.0380	0.1045	0.2390	0.3920	0.2120
10	0.0000	0.0000	0.0000	0.0005	0.0010	0.0045	0.0105	0.0425	0.2115	0.7295

Table 3A: Ex-Post Conditional Distributions for the NLSY/1979 (College Earnings Conditional on High School Earnings)
$Pr(d_i<Yc<d_i+1 \mid d_j<Yh<d_j+1)$ where d_i is the ith decile of the College Lifetime Ex-Ante Earnings Distribution and d_j is the jth decile of the High School Ex-Ante Lifetime Earnings Distribution
Information Set=$\{\theta_1,\theta_2,\theta_3,\theta_4,\theta_5,\theta_6\}$
Corrrelation(Y_C,Y_H) = 0.2842

	College									
High School	1	2	3	4	5	6	7	8	9	10
1	0.2118	0.1614	0.1188	0.0932	0.0782	0.0654	0.0532	0.0554	0.0651	0.0974
2	0.1684	0.1777	0.1557	0.1213	0.1038	0.0862	0.0640	0.0516	0.0417	0.0296
3	0.1374	0.1676	0.1464	0.1390	0.1244	0.0954	0.0754	0.0577	0.0333	0.0234
4	0.1080	0.1336	0.1433	0.1378	0.1213	0.1115	0.0980	0.0746	0.0475	0.0243
5	0.0787	0.1105	0.1232	0.1335	0.1345	0.1291	0.1144	0.0862	0.0614	0.0286
6	0.0656	0.1028	0.1149	0.1201	0.1276	0.1330	0.1250	0.0998	0.0823	0.0288
7	0.0548	0.0779	0.0842	0.1097	0.1196	0.1224	0.1410	0.1331	0.1132	0.0441
8	0.0428	0.0507	0.0741	0.0880	0.0994	0.1224	0.1410	0.1585	0.1539	0.0693
9	0.0416	0.0436	0.0474	0.0577	0.0803	0.1001	0.1277	0.1728	0.1939	0.1348
10	0.0386	0.0204	0.0269	0.0292	0.0339	0.0520	0.0704	0.1155	0.1945	0.4186

Table 3B: Ex-Post Conditional Distributions for the NLS/1966 (College Earnings Conditional on High School Earnings)
$Pr(d_i<Yc<d_i+1 \mid d_j<Yh<d_j+1)$ where d_i is the ith decile of the College Lifetime Ex-Ante Earnings Distribution and d_j is the jth decile of the High School Ex-Ante Lifetime Earnings Distribution
Information Set=$\{\theta_1,\theta_2,\theta_3,\theta_4,\theta_5\}$
Corrrelation(Y_C,Y_H) = 0.6226

	College									
High School	1	2	3	4	5	6	7	8	9	10
1	0.4001	0.1813	0.1023	0.0717	0.0611	0.0406	0.0422	0.0306	0.0337	0.0364
2	0.2144	0.2239	0.1663	0.1207	0.0862	0.0676	0.0486	0.0261	0.0256	0.0205
3	0.1286	0.1716	0.1591	0.1496	0.1181	0.0960	0.0695	0.0515	0.0340	0.0220
4	0.0870	0.1426	0.1551	0.1576	0.1386	0.1131	0.0810	0.0650	0.0365	0.0235
5	0.0450	0.0905	0.1390	0.1400	0.1405	0.1395	0.1165	0.0960	0.0625	0.0305
6	0.0350	0.0720	0.1126	0.1196	0.1456	0.1416	0.1306	0.1211	0.0900	0.0320
7	0.0210	0.0600	0.0710	0.1046	0.1201	0.1521	0.1466	0.1531	0.1126	0.0590
8	0.0205	0.0320	0.0455	0.0816	0.0951	0.1261	0.1562	0.1797	0.1667	0.0966
9	0.0180	0.0205	0.0305	0.0430	0.0755	0.0830	0.1476	0.1741	0.2316	0.1761
10	0.0125	0.0115	0.0235	0.0135	0.0225	0.0415	0.0611	0.1041	0.2077	0.5020

Table 4

Mean Rates of Return to College by Schooling Group

	NLS/1966		NLSY/1979	
Schooling Group	Mean Returns	Standard Error	Mean Returns	Standard Error
High School Graduates	0.2937	0.0083	0.3095	0.0113
College Graduates	0.3107	0.0114	0.3994	0.0129
Individuals at the Margin	0.3081	0.0446	0.3511	0.0535

Table 5

Percentage that Regret Schooling Choices

Schooling Group	NLS/1966	NLSY/1979
Percentage of High School Graduates who Regret Not Graduating from College	0.0966	0.0749
Percentage of College Graduates who Regret Graduating from College	0.0337	0.0311

Table 6A
Evolution of Uncertainty

Panel A: NLS/1966

	College	High School	Returns
Total Residual Variance	460.63	284.81	351.40
Variance of Unforecastable Components	181.37	128.43	327.35

Panel B: NLSY/1979

	College	High School	Returns
Total Residual Variance	709.75	507.29	906.01
Variance of Unforecastable Components	372.35	272.36	432.87

Panel C: Percentage Increase

	College	High School	Returns
Percentage Increase in Total Residual Variance	54.08%	78.12%	157.83%
Percentage Increase in Variance of Unforecastable Components	105.30%	112.07%	32.24%

Panel D: Percentage Increase in Total Variance due to Increase in Variance of Uncertainty

	76.66%	64.69%	19.03%

Table 6B

Evolution of Heterogeneity

Panel A: NLS/1966

	College	High School	Returns
Total Residual Variance	460.63	284.81	351.40
Variance of Forecastable Components (Heterogeneity)	279.25	156.38	24.05

Panel B: NLSY/1979

	College	High School	Returns
Total Residual Variance	709.75	507.29	906.01
Variance of Forecastable Components (Heterogeneity)	337.40	234.93	473.13

Panel C: Percentage Increase

	College	High School	Returns
Percentage Increase in Total Residual Variance	54.08%	78.12%	157.83%
Percentage Increase in Variance of Forecastable Components	20.82%	50.23%	1866.91%

Table 7

Share of Variance of Business Cycle in Total Variance of Unforecastable Components

	NLS/1966		NLSY/1979	
	Point Estimate	Standard Error	Point Estimate	Standard Error
High School	0.0586	0.0060	0.0069	0.0009
College	0.1193	0.0126	0.0158	0.0021

Let $Y_{s,t}$ denote the labor income in schooling sector s at age t. Let d_k denote the cohort dummy that takes the value one if the agent was born in year k and zero otherwise. Let X denote the vector of variables containing a dummy indicating whether the agent lived in the South Region at age 14 and a constant term. Let θ_j denote the factor j and $\alpha_{s,t,j}$ denote its factor loading at schooling sector s and age t. Let $\varepsilon_{s,t}$ denote the uniqueness. The model is:

$$Y_{s,t} = X\beta_{s,t} + \sum_{k=\tau_0}^{\tau_1} \gamma_{k,s,t} d_k + \theta_1 \alpha_{s,t,1} + \theta_2 \alpha_{s,t,2} + \theta_3 \alpha_{s,t,3} + \theta_4 \alpha_{s,t,4} + \theta_5 \alpha_{s,t,5} + \theta_6 \alpha_{s,t,6} + \varepsilon_{s,t}.$$

The cohort dummies can capture aggregate shocks. Under this interpretation, we test and reject the hypothesis that the agents know the aggregate shocks at the time of the schooling choice. We test and reject the hypothesis that the agent knows the uniqueness $\varepsilon_{s,t}$ and factors $\theta_{4,}\theta_5$, and θ_6 at the time of the schooling choice. Consequently, the total unforecastable component (aggregate and idiosyncratic components) is given by:

$$\tilde{P}_{s,t} = \sum_{k=\tau_0}^{\tau_1} \gamma_{k,s,t} d_k + \theta_4 \alpha_{s,t,4} + \theta_5 \alpha_{s,t,5} + \theta_6 \alpha_{s,t,6} + \varepsilon_{s,t}.$$

In school sector s lifetime earnings, this component is given by the discounted summation:

$$\tilde{Q}_s = \sum_{t=22}^{41} \left[\frac{\sum_{k=\tau_0}^{\tau_1} \gamma_{k,s,t} d_k}{(1+\rho)^{t-22}} \right] + \sum_{t=22}^{41} \left[\frac{\theta_4 \alpha_{s,t,4} + \theta_5 \alpha_{s,t,5} + \theta_6 \alpha_{s,t,6} + \varepsilon_{s,t}}{(1+\rho)^{t-22}} \right].$$

The variance of the total unforecastable component (aggregate plus idiosyncratic uncertainty) is:

$$Var\left(\tilde{Q}_s\right) = Var\left(\sum_{t=22}^{41} \left[\frac{\sum_{k=\tau_0}^{\tau_1} \gamma_{k,s,t} d_k}{(1+\rho)^{t-22}} \right] \right) + Var\left(\sum_{t=22}^{41} \left[\frac{\theta_4 \alpha_{s,t,4} + \theta_5 \alpha_{s,t,5} + \theta_6 \alpha_{s,t,6} + \varepsilon_{s,t}}{(1+\rho)^{t-22}} \right] \right).$$

The share of aggregate uncertainty in the total variance of the unforecastable component, m_s, is:

$$m_s = \frac{Var\left(\sum_{t=22}^{41} \left[\frac{\sum_{k=\tau_0}^{\tau_1} \gamma_{k,s,t} d_k}{(1+\rho)^{t-22}} \right] \right)}{Var\left(\tilde{Q}_s\right)}.$$

In the table, we plot m_s for s = high school, college, for both the NLSY/1979 and NLS/1966. For example, 5.86% of the total variance of unforecastable components in high-school lifetime earnings is due to the aggregate uncertainty in the NLS/1966 sample and 0.7% in the NLSY/1979 sample.