working paper

11 25

Ability Matching and Occupational Choice

Jonathan James

FEDERAL RESERVE BANK OF CLEVELAND

Working Paper 11-25 October 2011

Ability Matching and Occupational Choice

Jonathan James

This paper develops and estimates an individual model of occupational choice and learning that allows for correlated learning across occupation-specific abilities. As an individual learns about their occupation-specific ability in one occupation, this experience will be broadly informative about their abilities in all occupations. Workers continually process their entire history of information, which they use to determine when to change careers, as well as which new career to go to. Endogenizing information in this manner has been computationally prohibitive in the past. I estimate the model in an innovative way using the Expectation and Maximization (EM) algorithm. The model is estimated on the National Longitudinal Survey of Youth 1997. The estimates suggest that both direct and indirect learning play an important role in early career wage growth, with those with the lowest levels of education achieving the largest increases.

JEL codes: J24, J31, J62

Keywords: Occupational choice, correlated learning, dynamic discrete choice, expectation and maximization algorithm.

Jonathan James is at the Federal Reserve Bank of Cleveland. He can be reached at jwj8@duke.edu. He is grateful to Peter Arcidiacono, Joe Hotz, and Pat Bayer for providing advice, comments, and encouragement. He extends special thanks to seminar participants at Duke University, UNC-Chapel Hill, the University of Pittsburgh, and the Federal Reserve Bank of Cleveland for helpful comments and feedback.

1 Introduction

A salient feature of the labor market is the high degree of occupational mobility of younger workers. Kambourov and Manovskii (2008) find that male workers, ages 23-28 with at most a high school degree have a 30% chance of changing occupations in a given year, even when occupation is defined using a very broad level of aggregation (the census 1-digit level). The leading models that address this feature of the labor market are the occupational matching models of Miller (1984), Neal (1999), and Pavan (2009). These models, along with the job shopping literature of Johnson (1978), Jovanovic (1979), and Topel and Ward (1992), argue that this high frequency of occupational change is the result of an experimentation and learning process that younger workers undergo to search across different types of employment to find their comparative advantage.

These matching models have provided a powerful framework for understanding early career occupational mobility and wage growth. However, with few exceptions (e.g. Shaw (1987)), this literature has made the strong assumption that the nature of occupational match is exclusively occupation specific. This independence assumption implies that as an individual learns about their ability in one occupation, this experience provides no additional information about their ability in any other occupation. While this simplifying assumption eases the computational burden of these models, it unfortunately severely limits how much we can learn about occupational sorting and mobility and it's implications for worker welfare.[1]

This paper develops and estimates an occupational matching model that relaxes this assumption of independence. Namely, occupational matches will share a robust correlation structure, such that as a worker learns about their ability in one occupation, it will be informative about their abilities in all occupations. Relaxing this assumption fundamentally changes the worker's problem, greatly broadening the model's usefulness and our understanding of why workers change occupations.

The focus of the previous literature has been on occupational mobility due to learning about a

[1] Learning models outside of labor economics also have assumed independence. For example, Crawford and Shum (2005) evaluate a pharmaceutical matching model in which patients are uncertain about their innate responsiveness to Anti-ulcer medications. The patients experience with their current medication does not affect their choice of subsequent medications. Likewise, Ackerberg (2003) assumes independence in an advertising and learning model over yogurt products. Allowing for correlation in his model modifies the consumers problem in a number of interesting ways.

poor match. An important contribution of the paper is that by relaxing the independence assumption, we are able to generalize the role of learning to account for many new patterns of mobility that we observe in the data. Specifically, the model will be able to additionally rationalize 1) occupational changes due to promotion, where high ability workers change occupations precisely because they are high ability, 2) occupational clustering, where workers may tend to switch occupations between a handful of apparently similar occupations (e.g. between business operations occupations and professional occupations), and 3) occupational cycling, where workers leave an occupation, only to return a few periods later.[2] In more general terms, the correlated learning modifies the workers problem in an ingenious way so that learning now has an important two dimensional role of affecting the decision to change occupations as well as driving the decision of which occupation to go to next.[3] Generalizing the learning framework to account for these other mobility patterns is an important step for understanding more thoroughly how workers substitute across different types of employment.

The second contribution of the paper is that it extends the occupational matching literature by simultaneously endogenizing the education decision along with occupational choice and learning. The previous literature abstracts from the education decision and estimates models on individuals with homogenous education. Miller (1984) argues that the data indicates that college graduates appear to make more informed decisions about occupational matches than high school graduates. The ability matching model allows individuals to have heterogeneous ability for schooling. In an intuitive way, through the occupational correlation structure, the model allows schooling ability to be informative about prospective occupational matches. Endogenizing education in this manner will provide important insight into the relationship between educational attainment, the occupational learning process, and the source of information differentials among education groups.

The final contribution of the paper is methodological. Allowing occupation specific ability to be correlated across occupations comes at a huge computational cost. First, the number of

[2]Kambourov and Manovskii (2008) show that about 30% of occupational changers actually return to their previous 1-digit occupation within four years.

[3]The independence assumption allows workers to leave occupations because of poor matches, but it does not explain how they choose new occupations because independence means workers view all occupations as *ex-ante* identical.

parameters are increasing in the number of occupations, with which traditional approaches require jointly maximizing a large set of parameters. Second, since abilities are unobserved, we are required to integrate this high dimensional correlated unobservable out of the data. Finally, the size of the state space is overwhelming, containing all potential beliefs about occupational abilities, such that solving the dynamic model is extremely difficult.

These three features independently are challenging but combined make traditional methods computationally infeasible. I address these challenges by estimating the model using the expectation and maximization (EM) algorithm. The EM algorithm greatly simplifies the problem by allowing many of the model parameters to be estimated separately, avoiding joint maximization. Furthermore, it eliminates the need to perform the high dimensional correlated integration and replaces it with a set of single dimensional integrals, which conveniently admit a closed form expression. Perhaps most important, is that by utilizing the EM algorithm, the computational burden only increases linearly in the number of occupations (as oppose to exponentially with traditional methods). This allows the empirical model to accommodate a large number of occupations avoiding coarsely aggregating occupations into a smaller set.[4] Finally, as shown in Arcidiacono and Miller (2010), the EM algorithm provides a means to recover consistent estimates of conditional choice probability (CCP's) with unobserved state variables, which can be used to tractably estimate the dynamic structural choice model.

I estimate the model using the 1997 cohort of the National Longitudinal Survey of Youth (NLSY97). The empirical model allows for a large choice set where individuals choose among ten occupations. The NLSY97 data is invaluable for a model that requires a richer detail of occupations. A highly documented issue with its counterparts is the prevalence of spurious occupational changes. The data issues are largely due to how the data is gathered. The occupation instruments in the NLSY97 data are collected in such a way to reasonably avoid these measurement issues.

[4] Avoiding coarse aggregation of occupations (e.g. blue collar and white collar) is highly desirable in the context of the learning model. Aggregation assumes that human capital (i.e. schooling, experience, and occupational ability) is perfectly transferable among the grouped occupations, which if untrue will bias many of the primitives of the learning model. For example, if an individual learns they have a bad match as a restaurant waiter and changes careers to become a welder, which they learn they have a high ability in, then when these occupations are aggregated to blue collar, the model will interpret these wage increases as returns to tenure when they are actually returns to search. A more thorough discussion of misspecification bias is presented in appendix A.

The estimates on the distribution of abilities suggests that workers can potentially achieve large wage increases by finding occupations they are well suited for. Furthermore, the estimates of the covariance implies that correlation is important to the learning process. In addition, the correlation between schooling ability and the occupational abilities suggest that as workers learn about their schooling ability, this generates very different beliefs about expected abilities across occupations.

Using the parameter estimates, I simulate careers to further investigate the role of sorting on ability by educational group and it's influence on wage growth. The simulations show that high school dropouts and high school graduates are more likely to change occupations and more likely to sort on ability than college graduates. Consequently, workers without a college degree achieve the greatest increases in wages due to sorting, where within the initial years of their careers (8-9 years labor market experience), sorting on ability attributes to an increase of wages of 9.5% for high school drops and 7.5% for high school graduates, compared to 5.2% for college graduates. Previous work (e.g. Miller (1984)) concluded that more informed career decisions resulted in the infrequent occupational mobility of college graduates. However, the results from the simulations suggest a different interpretation, which is that college graduates are less likely to change occupations because sorting on ability is less important to these workers and occupational changes are driven by other factors. For example, a college graduate may be more likely to stay in a poorly matched occupation because the returns to their college degree are higher in that occupation.

A number of recent papers have aimed at allowing for correlated learning in empirical work. Antonovics and Golan (2011); Sanders (2010) allow for correlation in occupational learning models. There method utilizes information in the Dictionary of Occupational Titles to construct a measure of manual and cognitive skills for each occupation following Yamaguchi (2010). The correlation matrix is constructed from these two measured factors and learning is reduced to these two dimensions. The empirical approach in this paper is to have each occupation represent it's own factor, where the correlated learning is over J factors (the number of occupations in the economy). In addition, the correlation among these J factors is estimated rather than drawn from auxiliary data. Dickstein (2011) uses a different approach to allow for correlated learning among anti-depressant medications. He incorporates correlation into the dynamic allocation index rule used in Miller (1984). One drawback of this approach is that any additional factors (observed or unobserved) cannot affect the dynamic decision process because the optimal choice is based solely on learning parameters, abstracting from anything else in the utility function (in particular any unobserved random utility shocks).

The outline of the paper is as follows: In section 2, I discuss the details of the occupational choice model with ability matching. Section 3 outlines the two stage estimation strategy using the EM algorithm that uncovers the structural parameters. Section 4 will discuss the population extracted from the NLSY97 data. Section 5 reveals the empirical results. Section 6 simulates data from the parameter estimates to investigate the relationship between sorting on ability and wage growth with an emphasis on it's link to educational attainment. Finally section 7 concludes.

2 Ability Matching Model of Occupation Choice

This section outlines the dynamic career decision model. The essential feature of the model is that workers vary in their innate skills and abilities for each occupation in the economy and most importantly, these unique occupation specific abilities are unknown to the worker when they begin their career. Workers are able to learn about their occupation specific abilities through direct experience. In addition as workers are learning about their ability in one occupation, they will be able to make inferences about their abilities in all occupations by exploiting a known correlation structure between occupational abilities. This section first describes the occupational ability match vector and how it relates to wages, as well as the learning process. Then, it details the dynamic occupational choice model.

2.1 Ability Matches

The economy consists of J distinct (but potentially related) occupations. Workers have an innate ability for each of the J occupations. We can uniquely characterize an individual by their vector of ability endowments,

$$\mathcal{M}_i = [\quad \mu_{i1} \quad , \quad \mu_{i2} \quad , \cdots , \quad \mu_{iJ} \quad , \quad \mu_{is} \quad]' \tag{1}$$

where μ_{i1} represents individual i's ability in occupation 1, with μ_{i2} being their ability in occupation 2, etc. Two important features of the ability endowment drive career decisions. First, the ability vector is initially unknown to the individual, and second, the components of the ability endowment are not independent.

The final term in the ability vector, μ_{is}, represents individual i's ability or preference for school. Similar to Belzil and Hansen (2002); Keane and Wolpin (1997) this is a heterogenous term which affects the utility for attending school in the form of additional cost.[5] This component will play a prominent role in the workers problem because, like the occupational abilities, school ability is not independent.

The ability match vector is distributed through the population multivariate normal, parameterized as,

$$\mathcal{M}_i \sim \mathcal{N}(\mathbf{E}(\mathcal{M}_i), \mathbf{V}(\mathcal{M}_i)) \tag{2}$$
$$\mathcal{N}\left(\begin{bmatrix} \Gamma_{(J\text{x}1)} \\ \hline \gamma_{s_{(1\text{x}1)}} \end{bmatrix}_{(J+1\text{x}1)}, \begin{bmatrix} \Delta_{(J\text{x}J)} & \delta'_{sJ_{(1\text{x}J)}} \\ \hline \delta_{sJ_{(J\text{x}1)}} & \delta^2_{s_{(1\text{x}1)}} \end{bmatrix}_{(J+1\text{x}J+1)}\right)$$

where Γ is a vector containing the mean of occupational abilities in the population. γ_s represents the mean of schooling ability. The covariance is the partition of the three terms, Δ, δ_s^2, and δ_{sJ}. Δ is the occupation ability covariance matrix. It's dimension is JxJ. δ_s^2 is the variance of schooling ability, and δ_{sJ} is the covariance between schooling ability and each of the occupational abilities.

If workers have rational expectations, this covariance structure is going to play a critical role in the workers learning process. Specifically, workers will not only learn about their ability in the occupations they have direct experience in, but they will also be able to exploit this covariance structure to make inferences about their abilities in other occupations. Likewise, if schooling ability is correlated with any of the occupational abilities, as workers learn about their schooling ability, this will also provide valuable information about their abilities in the various occupations.

The covariance structure is an important element in the model because it's sparseness (dense-

[5]The model is agnostic about what this term actually represents, however for consistency it will be referred to as schooling ability (where it could also be referred to as schooling match, schooling preference, or schooling cost).

ness) represents the importance of comparative (absolute) advantage. In one case, if abilities are highly correlated, then efficient search becomes less important since the individual's problem effectively reduces to a single dimension. Any path of experience will effectively yield the same information, where they are discovering whether they are a high type or a low type. In the other extreme case, if abilities are only loosely correlated, then efficient search becomes a critical part of the career process as workers seek to find their comparative advantage. In this case, the experience path matters. As workers gain labor market experience, it becomes more costly for them to experiment with new careers because this likely results in the loss of accumulated human capital in their current occupation. The fact that searching later in the career is more costly places more importance on strategically finding a good occupational match early in the career.

Going forward, it's assumed that when individuals begin making decisions in the model, their schooling ability is known.[6] Knowing their schooling ability, μ_{is}, workers can use the distributional parameters to form updated beliefs about their occupational abilities. Letting $\{\mathbf{E}_{it}(\mathcal{M}_i), \mathbf{V}_{it}(\mathcal{M}_i)\}$ indicate the mean and variance, which completely characterizes their beliefs about their ability vector given information up to date t, then their beliefs at $(t = 1)$ are described by,

$$\begin{aligned}
\{\mathbf{E}_{i1}(\mathcal{M}_i), \mathbf{V}_{i1}(\mathcal{M}_i)\} &= \{\mathbf{E}(\mathcal{M}_i), \mathbf{V}(\mathcal{M}_i) | \mu_{is}\} \\
&= \left\{ \mathcal{N}\left(\left[\begin{array}{c} \Gamma \\ \hline \gamma_s \end{array} \right], \left[\begin{array}{c|c} \Delta & \delta_{sJ}^{'} \\ \hline \delta_{sJ} & \delta_s^2 \end{array} \right] \right) \middle| \mu_{is} \right\} \\
&= \{\Gamma + \delta_{sJ}(\delta_s^2)^{-1}(\mu_{is} - \gamma_s), \Delta - \delta_{sJ}(\delta_s^2)^{-1}\delta_{sJ}^{'}\}
\end{aligned} \tag{3}$$

where the last line of equation (3) is the conditional distribution of a multivariate normal, whose resulting distribution is also multivariate normal.

This generates an important degree of heterogeneity across individuals in their beliefs about their occupation specific abilities. Likewise, it creates an important link between education decisions and occupational search patterns.

[6] This may be due to the fact that they either actually know this value, or they have learned it through a sufficient number of years of school.

2.2 Wages and Learning

Workers possess different types of accumulated human capital (e.g. education and work experience). Each occupation utilizes these forms of human capital in a unique way, so the individual's productivity will vary by occupation. In a similar way, their occupation specific ability will also affect productivity. Following the literature on productivity wages, if a worker is employed in occupation $j \in J$, they are paid their occupation specific marginal product with log-wage described by.

$$w(h_{it}, \mu_{ij}, \eta_{ijt}) = h_{it}\theta_j + \mu_{ij} + \eta_{ijt} \tag{4}$$

where i subscripts the individual, t the time period, and j the occupation. The term, h_{it}, represents the individual's vector of accumulated human capital, including experience in the various occupations and education at date t. θ_j is the occupation specific return for the different types of accumulated human capital. Their occupation specific ability is represented by μ_{ij}. The term η_{ijt} is an individual, occupation specific, transitory technology shock to wages.

Since occupational ability is additive in the log-wage equation, it has a very clear interpretation as a percentage difference in wages, such that an individual with a value of $\mu_{ij} = 0.1$ enjoys a 10% higher wage in occupation j, all else equal, while an individual with $\mu_{ij} = -0.1$ has 10% lower wages.

When an individual works in occupation j, they observe their wage. Following Miller (1984) and Jovanovic (1979), what they do not observe is the individual values of the random variables, μ_{ij} and η_{ijt}. They can only observe their sum by differencing $(w_{ijt} - h_{it}\theta_j) = (\mu_{ij} + \eta_{ijt})$. Workers are extremely interested in knowing their occupational ability because it is a potentially large (or negative), persistent component affecting future returns in that occupation. Given this noisy signal, individuals make inferences about their abilities via Bayes' rule, using their priors and knowledge of the distribution of the technology shock,

The technology shock η_{ijt} represents the noise of the ability signal. This random variable has an occupation specific variance that is normally distributed and *i.i.d.* across time and individuals.

$$\eta_{ijt} \sim \mathcal{N}(0, \sigma_j^2) \quad \forall \quad j \in \{1, 2, \cdots, J\}$$

Occupations differ in the variance of the technology shock, implying that productivity in some occupations may be more or less influenced by transitory factors. In occupations with low σ_j^2, the individual will be able to infer their occupation ability match faster than occupations with a high transitory noise. Taken to the extreme, if $\sigma_j^2 = 0$, then the worker will be able to infer their ability from a single wage observation. From the worker's perspective, one dimension occupations can be classified is along fast learning and slow learning occupations, which is dictated by σ_j^2.

A second dimension in which occupations differ is how informative their abilities are about the abilities in other occupations. Workers distinguish between high information occupations and low information occupations. High information occupations are not only informative about the own occupation ability but also provide information about abilities across occupations. Low information occupations are those that only provide information about the own occupation ability.

In the model, workers only observe their productivity (wage) after they choose an occupation.[7] This means that when workers make career decisions they choose occupations based on expected wages, taking into account what they have learned so far in their career. The details of the decision process are outlined in section 2.3, the emphasis here is to highlight that workers are precluded from receiving multiple wage signals simultaneously and must incur the cost of accepting a job in order to acquire information. If this were not the case, workers observe all of their productivities for each occupation in each period, in which learning would simply be a product of time and individuals require no strategy to efficiently seek out information.

When a worker observes an ability signal they update their beliefs recursively using Bayes's rule. Their posterior beliefs are a function of their initial beliefs, the noisy signal of their ability , and the known distribution of the technology shock corresponding to the occupation they worked. Degroot (1970) outlines the updating formulas given a normally distributed prior and normally distributed

[7] Other models of occupational choice like Keane and Wolpin (1997); Sullivan (2010) assume individuals observe all of their potential wages for each occupation in each period prior to making a choice.

signal. The updated beliefs can be described in matrix form using the following notation. Let $d_{it-1} = j$ indicate that individual i worked in occupation j in period $t-1$. Then,

$$\zeta_{it-1_{J\mathrm{x}1}} = \begin{cases} \zeta_{it-1_j} = w_{ijt-1} - h_{it-1}\theta_j \quad , & \mathit{if}\ (d_{it-1} = j) \\ \qquad\qquad 0 \quad , & \mathit{otherwise} \end{cases}$$

$$\Sigma_{it-1_{J\mathrm{x}J}} = \begin{cases} \Sigma_{it-1_{j,j}} = \dfrac{1}{\sigma_j^2} \quad , & \mathit{if}\ (d_{it-1} = j) \\ \qquad\qquad 0 \quad , & \mathit{otherwise} \end{cases}$$

$\zeta_{it-1_{J\mathrm{x}1}}$ is a sparse vector that contains all zeros except in the j^{th} position, which has the ability signal of occupation j. $\Sigma_{it-1_{J\mathrm{x}J}}$ is a sparse matrix with zeros everywhere except in the (j,j) element, which contains the inverse of the occupation j technology shock variance.

Given this vector and matrix, using the previous periods beliefs about occupation ability, which are multivariate normal described by their mean $\mathbf{E}_{it-1}(\mathcal{M}_i)$ and covariance $\mathbf{V}_{it-1}(\mathcal{M}_i)$, the posterior beliefs are multivariate normal with mean and variance defined as,

$$\begin{aligned} \mathbf{E}_{it}(\mathcal{M}_i) &= \Big[\mathbf{V}_{it-1}(\mathcal{M}_i)^{-1} + \Sigma_{it-1}\Big]^{-1}\Big[\mathbf{V}_{it-1}(\mathcal{M}_i)^{-1}\mathbf{E}_{it-1}(\mathcal{M}_i) + \Sigma_{it-1}\zeta_{it-1}\Big] \\ \mathbf{V}_{it}(\mathcal{M}_i) &= \Big[\mathbf{V}_{it-1}(\mathcal{M}_i)^{-1} + \Sigma_{it-1}\Big]^{-1} \end{aligned} \tag{5}$$

The previous periods beliefs take into account the entire past sequence of outcomes and choices as well as their initial priors, therefore it is sufficient to only condition on the previous periods beliefs, rather than the entire history. Also, equation (5) demonstrates that if an individual receives no signal (does not work) then the update simply returns the prior.

2.3 A Dynamic Model of Career Decisions with Learning

The dynamic occupational choice model shares a similar framework to other structural occupational choice models like Keane and Wolpin (1997) and Sullivan (2010). The distinguishing difference is that individuals now face uncertainty over their occupation specific abilities, and given their forward looking behavior, workers seek an optimal career path to efficiently learn about their unknown abilities.

From age 16 ($t = 1$) until age T, individuals make career decisions to maximize expected lifetime utility. In each period, individuals make a discrete decision $d_{it} \in \{\mathrm{u}, \mathrm{s}, 1, 2, \cdots, J\}$ among

$J+2$ mutually exclusive alternatives, where $d = \text{u}$ is unemployment (home production), $d = \text{s}$ is attending school, and $d = j$, for each $j \in \{1, 2, \cdots, J\}$, is employment in occupation j. Each choice is associated with a current period utility $u_d(\cdot)$ and a non-pecuniary random utility shock ε_d, which is additive and *i.i.d.* across choices and time, so the per-period pay-off for choice d is, $u_d(\cdot) + \varepsilon_d$

2.3.1 Employment: $d \in \{1, 2, \cdots, J\}$

The per-period utility associated with employment in occupation j is,

$$
\begin{aligned}
u_{ijt}(\cdot) = \alpha_{j1} + \alpha_{j2}(d_{it-1} \neq j) &+ \alpha_{j3}(Expr_j = 0) \\
&+ \alpha_w \mathbb{E}\Big(u(\exp(w_{ijt})) \Big| h_{it}, \mathbf{E}_{it}, \mathbf{V}_{it} \Big) \\
&\forall j \in \{1, 2, \cdots, J\} \\
\text{where, } u(\exp(w)) = \frac{1}{1-\rho} \exp(w)^{(1-\rho)} \quad &, \rho \in [0, \infty)
\end{aligned}
\tag{6}
$$

This expression contains a constant, α_{j1}, which represents the non-pecuniary benefits of being employed in occupation j. An entry cost, α_{j2}, if the worker did not work in occupation j in the previous period, and α_{j3}, an additional entry cost if the worker has never worked in occupation j. Each parameter is occupation specific except for the parameter on expected utility of wages, α_w, which is the same regardless of the source of the wage.[8]

An important departure from previous work is that workers are not assumed to be maximizing expected lifetime wages. Rather, because workers face a high degree of uncertainty in this model, it is likely that risk plays an important role in career decisions. Assuming workers cannot save, they put all of their wages into consumption. A risk averse worker may be unwilling to try their hand in a new occupation in which their is a good chance they may receive a very low wage and thus lower consumption. The iso-elastic utility specification in equation (6) allows risk over wages to influence workers' decisions. Risk aversion is measured by the constant relative risk aversion parameter ρ. If $\rho = 0$, then workers are expected wage maximizers.

[8]Recall that equation (6) has $\exp(w)$ as the argument because w represents log wage.

Following the expression from equation (4), log wages are given by,

$$
\begin{aligned}
w(h_{it}, \mu_{ij}, \eta_{ijt}) &= h_{it}\theta_j + \mu_{ij} + \eta_{ijt} \\
=&\theta_{j1} + \theta_{j2}(ed) + \theta_{j3}(ed^2/100) + \theta_{j4}(Expr_j > 0) \\
&+ \theta_{j5}(Expr_j) + \theta_{j6}(Expr_j^2/100) + \theta_{j7}(\sum_{-j} Expr_k) + \mu_{ij} + \eta_{ijt}
\end{aligned} \tag{7}
$$

where ed is the number of years of education, $Expr_j$ is the total labor market experience in occupation j, and $\sum_{-j} Expr_k$ is the sum of other labor market experience excluding j.

Deriving the expression for the expected utility of wages for occupation j using the expression in equation (4) yields,

$$
\begin{aligned}
&\mathbb{E}\Big(u(\exp(w_{ijt}))\Big| h_{it}, \mathbf{E}_{it}, \mathbf{V}_{it}\Big) \\
=&\mathbb{E}\Big(\frac{1}{1-\rho}\exp(w_{ijt})^{(1-\rho)}\Big| h_{it}, \mathbf{E}_{it}, \mathbf{V}_{it}\Big) \\
=&\mathbb{E}\Big(\frac{1}{1-\rho}\exp\big((1-\rho)(h_{it}\theta_j + \mu_{ij} + \eta_{ijt})\big)\Big|, \mathbf{E}_{it}, \mathbf{V}_{it}\Big) \\
=&\frac{1}{1-\rho}\exp\Big((1-\rho)(h_{it}\theta_j + \mathbf{E}_{it}(\mu_{ij})) + \frac{(1-\rho)^2}{2}(\mathbf{V}_{it}(\mu_{ij}) + \sigma_j^2)\Big)
\end{aligned} \tag{8}
$$

where the last line follows from the fact that the sum of the two unknown random components, the occupation j ability, μ_{ij}, and the transitory productivity shock, η_{ijt}, are normally distributed, making wages distributed log-normal. The term $\mathbf{E}_{it}(\mu_{ij})$ in equation (8) represents the individual's marginal expectation about their ability in occupation j given all of the available information at time t and likewise $\mathbf{V}_{it}(\mu_{ij})$ represents the marginal variance.

Contemporaneously, risk aversion is an additional search friction similar to the other entry costs. However, risk aversion may also play a role intertemporally as occupations deliver very different consumption streams through the returns to accumulated human capital. If workers are risk averse, they may find occupations with initially steeper returns to tenure more attractive than those with returns that have a more gradual profile.

2.3.2 Schooling and Unemployment: $d \in \{\mathbf{s}, \mathbf{u}\}$

If the individual chooses schooling in period t they receive flow utility.

$$\begin{aligned} u_{i\mathrm{s}t}(\cdot) =& \alpha_{\mathrm{s}1} + \alpha_{\mathrm{s}2}(AttCOL) + \alpha_{\mathrm{s}3}(ReentryHS) \\ &+ \alpha_{\mathrm{s}4}(ReentryCOL) + \alpha_{\mathrm{s}5}\mu_{is} \end{aligned}$$

Schooling preference depends on $(AttCOL)$ which indicates whether the individual is attending college, $(ReentryHS)$ and $(ReentryCOL)$ represents the re-entry cost if the individual was not enrolled in school in the previous period. Finally, the utility of schooling will be affected by school ability μ_{is}.

The mean flow utility of unemployment is normalized to zero (i.e. $u_{i\mathrm{u}t} = 0$).

2.3.3 Value Functions

Let $\mathcal{S}_{it}$ represent individual i's state vector at time t. This includes all of the variables relevant to the individuals decision, which given the current specification is years of education, labor market experience for each occupation j, the previous periods decision, d_{it-1}, and their beliefs about their occupational ability, $\mathbf{E}_{it}(\mathcal{M}_i)$ and $\mathbf{V}_{it}(\mathcal{M}_i)$. Furthermore, let $\boldsymbol{\varepsilon}_{it}$ be the $J+2$ current period random utility shocks, $\varepsilon_{\mathrm{u,s},1,2,\cdots,J}$

Workers seek to maximize the expected discounted sum of future utility. Let $V_t(\mathcal{S}_{it}, \boldsymbol{\varepsilon}_{it})$ be the value function for a worker in a particular state at time t when they make optimal decisions. Given that time is finite, $V_{T+1}(\cdot) = 0$. With discount of future utility β, we can represent the value function with the following recursive Bellman equation,

$$V_t(\mathcal{S}_{it}, \boldsymbol{\varepsilon}_{it}) = \max_{\substack{c \in \{\mathrm{u,s} \\ ,1,\cdots,J\}}} \{u_c(\mathcal{S}_{it}) + \varepsilon_c + \beta \mathbb{E} V_{t+1}(\mathcal{S}_{it+1}, \boldsymbol{\varepsilon}_{it+1} | \mathcal{S}_{it}, d_{it} = c)\} \tag{9}$$

The expectation is taken over next periods utility shocks, $\boldsymbol{\varepsilon}$, as well as all possible realization of the occupational ability signals, ζ, if they chooses one of the employment options, which effects their beliefs in the next period. The other state variables, education and labor market experience evolve deterministically, but are endogenous to the choice.

Letting $v_{ct}(\cdot)$ represent the conditional value function, which is the mean pay-off for each choice, excluding the current periods error term defined as,

$$v_{ct}(\mathcal{S}_{it}) = u_c(\mathcal{S}_{it}) + \beta \mathbb{E} V_{t+1}(\mathcal{S}_{it+1}, \boldsymbol{\varepsilon}_{it+1} | \mathcal{S}_{it}, d_{it} = c)$$
$$\forall c \in \{\mathrm{u}, \mathrm{s}, 1, 2, \cdots, J\} \tag{10}$$

then the individual's optimal choice in period t, d^*_{it} is determined by.

$$d^*_{it} = \underset{c \in \{\mathrm{u},\mathrm{s},1,\cdots,J\}}{\operatorname{argmax}} \{v_{ct}(\mathcal{S}_{it}) + \varepsilon_{ict}\} \tag{11}$$

Dynamics play an important role in the individual's decision process. Since the information the individual receives is conditional on the choice, individuals strategically make career decisions which are likely to give them the best information. In Miller (1984) individuals have an incentive early on to try out occupations where the ability match distribution has a low mean, but high variance because this could potentially lead to very valuable information if the individual's talent happens to be in the upper tail of the distribution.[9] Consequently, embedded in this maximization problem is an optimal search strategy in which individuals strategically search over occupations to efficiently learn about their innate abilities.

3 Estimation

The ability matching model possess a number of features that make estimation challenging. These problems are rooted in the fact that many components in the model grow exponentially complex in the number of occupations. A popular method seen in the literature to address this curse of

[9] Most interesting is that relaxing the independence assumption actually generates a potentially new set of optimal career paths than found in the previous literature. Consider a worker choosing between two occupations. The distribution of the unknown match value for occupation A is characterized by a high variance and low mean, while the distribution for occupation B has a low variance and high mean. The optimal career path described in Miller (1984) is for workers to always initially sort into occupation A, the high variance occupation, with the goal being that this is the occupation with the greatest informational benefit. However, if the occupational matches are correlated, then the worker's optimal career path may be to initially sort into occupation B and through the correlation structure learn about their match value in occupation A.

dimensionality is to reduce the number of occupations by coarsely aggregating occupations into a smaller set. Unfortunately, in the context of the learning model this is a particularly dangerous approach because there is a serious risk of misspecification, which will bias nearly all of the model parameters.[10]

In order to estimate the model allowing for a rich set of occupations, we require a new approach which explicitly addresses the curse of dimensionality. The curse of dimensionality presents itself in a number of ways. First, given that we do not observe the high dimensional vector of occupational and schooling ability, we need to integrate over these correlated unobservables. With many occupations this high-dimensional integration becomes increasingly difficult. The second computational problem stems from the fact that nearly all of the parameters in the model are occupation specific. With many occupations, this corresponds to many parameters. The burden of jointly estimating all of these parameters simultaneously poses a major computational challenge, in particular when using numerical gradient based methods. Finally, the state space in the model is continuous along multiple dimensions as it includes all potential combinations of beliefs about occupational abilities. Even if beliefs where the only state variable and for each occupation, beliefs were crudely gridded to 10 points, then with 10 occupations the state space would contain over 10 billion combinations of beliefs. Traditional full solution methods for solving dynamic discrete choice problems will not be feasible.

The estimation strategy utilizes the Expectation and Maximization (EM) algorithm to overcome these computational challenges. This approach breaks the curse of dimensionality so that the computational complexity only grows linearly in the number of occupations, allowing many occupations into the empirical model. Compared to traditional maximum likelihood, which attempts to solve this extremely complex problem all at once, the EM algorithm can be viewed as a process which breaks the larger (infeasible) problem up into a couple of smaller (feasible) problems, in particular an expectation step and a maximization step. By iterating between these two interlinked steps, Dempster et al. (1977) show that the parameters converge to the same solution as numerical maximum likelihood.[11]

[10] Misspecification bias and it's consequences are discussed in appendix A.
[11] A local maximum.

With respect to the computational challenges outlined above, the EM algorithm breaks the $J-$dimensional correlated integral into J single dimensional integrals. Furthermore the maximization becomes additively separable for each of the wage equations and the choice equation, allowing us to estimate each set of parameters separately. Finally, as demonstrated by Arcidiacono and Miller (2010), the EM algorithm provides a means to recover conditional choice probabilities (CCP's) from the data in the case of unobserved state variables. Hotz and Miller (1993); Arcidiacono and Miller (2010) show that CCP's can be used as an alternative representation of the expected future value term, enabling us to solve the complicated dynamic structural model without solving the optimization at every point in the state space.

The estimation is implemented in two stages. In the first state, I use the EM algorithm to recover all of the model parameters except the structural utility parameters. Rather than resolving the structural model at each iteration of the algorithm, I replace the structural choice probabilities with a flexibly specified dynamic reduced form. Then, using the parameter estimates from the first stage, I estimate the dynamic structural model in the second stage. This section is broken into three sub-sections. The first describes the likelihood function and the challenges associated with traditional maximum likelihood. The next sub-section describes the EM algorithm, including the first stage of estimation. Finally, using the first stage estimates, I outline the estimation of the dynamic structural model.

Normalizations

In addition to the standard normalizations associated with estimating discrete choice models[12], some additional parameters of the model are not identified. Specifically, given that the ability matches are random variables, the mean of occupation ability will not be separately identified from the constant in the wage equation and the mean of schooling ability will not be separately identified from the constant in the schooling utility. Therefore, the mean of the ability vectors, Γ and γ_s, are set to zero. In addition, since schooling match enters the utility for schooling as, $\alpha_{s5}\mu_{is}$, the scale of schooling ability, $\mathbb{E}(\mu_{is}^2)$, will not be identified separately from the coefficient, α_{s5}. Therefore, I

[12]Only differences in utility functions are identified, and the variance of the random utility shock is not identified separately from the scale of the utility parameters

normalize $\delta_s^2 = 1$.

3.1 The Likelihood Function

The likelihood function for individual i is characterized as the joint probability of the observed career decisions $d_{it} \in \{\mathrm{u}, \mathrm{s}, 1, 2, \cdots, J\}$ and realized wages, w_{it}, for $t = 1$ to T, taking expectations of the likelihood over the unobserved schooling and occupational ability, $\{\mu_{i1}, \mu_{i2}, \cdots, \mu_{iJ}, \mu_{is}\}$.

$$\begin{aligned}\mathcal{L}_i = \int_{\tilde{\mu}_s}\int_{\tilde{\mu}_1}\cdots\int_{\tilde{\mu}_J}\Bigg[\prod_{t=1}^{T}\Bigg[&\Pr(d_{it}|x_{it}, b_{it}, \tilde{\mu}_s, \Lambda, \Theta) \\ &\times \prod_{j=1}^{J}\Pr(w_{it}|x_{it}, \tilde{\mu}_j, \Theta)^{d_{it}=j}\Bigg] \\ &\Bigg] H(\tilde{\mu}_1, \cdots, \tilde{\mu}_J, \tilde{\mu}_s|\Delta, \delta_{sJ})\end{aligned} \tag{12}$$

In equation (12), x_{it} denotes the observed data, which effects choices and wages. b_{it} represents the unobserved beliefs the individual has about their occupational ability. The structural parameters are summarized by $\Lambda \in \{\alpha_1, \cdots, \alpha_J, \alpha_w, \rho, \alpha_{\mathrm{s}}\}$, and the wage parameters are summarized by $\Theta \in \{\theta_1, \cdots, \theta_J, \sigma_1, \cdots, \sigma_J\}$. The wage parameters enter the choice probabilities because expected wages are a component of the employment flow utilities.

With expressions for the probability of the observed choices and wages, we need to integrate out of the likelihood the unobserved ability vector, whose density is represented by $H(\cdot)$. Given the normality assumptions in the model and the normalizations stated previously, $H(\cdot)$ is parameterized as,

$$H(\tilde{\mu}_1, \cdots, \tilde{\mu}_J, \tilde{\mu}_s|\Delta, \delta_{sJ}) \sim \mathcal{N}\left(\mathbf{0}, \left[\begin{array}{c|c} \Delta & \delta_{sJ}' \\ \hline \delta_{sJ} & 1 \end{array}\right]\right) \tag{13}$$

A striking feature of the model is that the true occupational abilities do not directly enter into the probabilities of the observed choices. Instead, it is the beliefs about their occupational abilities, b_{it}, which drive choices. Intuitively what this means is that two individuals with identical beliefs about their occupational abilities will make the same choices in probability, even though their occupational abilities may be completely different. Therefore, once we condition the choice probabilities on the beliefs, we can pull this component outside of the multiple integration over the occupational abilities. In rewriting the likelihood, it is then helpful to break the density $H(\cdot)$ into the unconditional distribution of schooling, $G(\cdot)$, and a conditional (on schooling) distribution of occupational ability, $F(\cdot)$. Where,

$$G(\tilde{\mu}_s) \sim \mathcal{N}(0, 1)$$
$$F(\tilde{\mu}_1, \cdots, \tilde{\mu}_J | \tilde{\mu}_s, \Delta, \delta_{sJ}) \sim \mathcal{N}\left(\delta_{sJ}\tilde{\mu}_s, \Delta - \delta_{sJ}\delta_{sJ}^{'}\right)$$

The unobserved schooling ability plays an important role in the model. Not only does it dictate individual's willingness to attend school, but it also influences their initial expectations about their occupational abilities. Rather than assuming this important unobservable is randomly distributed across individuals, it is common practice in the literature to condition this unobservable on potentially relevant information outside of the model.[13] For estimation, unobserved schooling ability is determined by,

$$\mu_{is} = z_i\lambda + \nu_i \tag{14}$$
$$s.t.\ \mathbb{E}(\mu_{is}) = 0 \text{ and } \mathbb{E}(\mu_{is}^2) - \mathbb{E}(\mu_{is}) = 1$$

Where z_i is a vector of observed variables including highest grade completed by the mother and information about the highest grade completed by the individual at age 16.[14] λ are parameters to be estimated, and the residual ν represents the remaining unobserved components and is distributed $\mathcal{N}(0, \sigma_\nu^2)$. The equation is estimated with the normalizations as constraints. If the observed variables z_i have no relation to μ_{is} then all of the variance will be absorbed by the unexplained error,

[13] Keane and Wolpin (1997) allow the initial years of schooling at age 16 to affect the probability of the initial conditions.

[14] Specifically, z_i includes the highest grade completed of the mother, an indicator if the highest grade completed of the mother is missing, an indicator if the highest grade completed by the individual at age 16 is less than 9, and an indicator if the highest grade complete by the individual at age 16 is greater than 9.

ν, and $\sigma_\nu = 1$. Therefore, schooling ability is distributed,

$$G(\tilde{\mu}_s | z_i, \lambda, \sigma_\nu^2) \sim \mathcal{N}(z_i \lambda, \sigma_\nu^2) \tag{15}$$

Pulling the choice probabilities out of the integration of the occupational abilities and replacing the joint density $H(\cdot)$ with the conditional density $F(\cdot)$ for occupational abilities, and the density $G(\cdot)$ for schooling ability, the likelihood in equation (12) becomes,

$$\begin{aligned} \mathcal{L}_i = \int_{\tilde{\mu}_s} \Bigg[& \prod_{t=1}^{T} \Pr(d_{it} | x_{it}, b_{it}, \tilde{\mu}_s, \Lambda, \Theta) \\ & \times \int_{\tilde{\mu}_1} \cdots \int_{\tilde{\mu}_J} \left(\prod_{t=1}^{T} \prod_{j=1}^{J} \Pr(w_{it} | x_{it}, \tilde{\mu}_j, \Theta)^{d_{it}=j} \right) \\ & F(\tilde{\mu}_1, \cdots, \tilde{\mu}_J | \tilde{\mu}_s, \Delta, \delta_{sJ}) \Bigg] G(\tilde{\mu}_s | z_i, \lambda, \sigma_\nu^2) \end{aligned} \tag{16}$$

There are three issues which make directly maximizing the likelihood in equation (16) computationally infeasible. First is the integration of the high dimensional correlated unobservables. Second is the pervasiveness of the parameters. For example, the wage parameters enter into the wage probabilities, the choice probabilities, as well as the belief functions. Even more challenging, is the fact that the distributional parameters for the integration also enter in the choice probabilities since they are components of the belief functions. Analytical gradients in this environment are extremely difficult to calculate. This forces us to rely on numerical gradients for maximization with which such a large parameter space is extremely slow.

The final challenge to directly maximizing equation (16) is that at each guess of the parameters, we need to solve the individual's optimization problem in order to evaluate the probabilities of the observed choices. Not only does this require solving for all potential career paths, but it also requires solving the value function for all potential realizations of occupational ability. Even using non-full solution interpolation methods like Keane and Wolpin (1994), with large T and large J, the complexity of the state space with such a large number of continuous variables is overwhelming and even performing the backward recursion for one iteration of the maximization is extremely computationally intensive.

3.2 The EM Algorithm

In order to estimate the model, I first consider a simplified version of the likelihood in equation (16) specified as,

$$\begin{aligned} \mathcal{L}_i = \int_{\tilde{\mu}_s} \Bigg[\prod_{t=1}^{T} \prod_{c \in C} \Omega_c(x_{it}, \tilde{\mu}_s)^{d_{it}=c} & \qquad (17) \\ \int_{\tilde{\mu}_1} \cdots \int_{\tilde{\mu}_J} \left(\prod_{t=1}^{T} \prod_{j=1}^{J} \Pr(w_{it}|x_{it}, \tilde{\mu}_j, \Theta)^{d_{it}=j} \right) & \\ F(\tilde{\mu}_1, \cdots, \tilde{\mu}_J | \tilde{\mu}_s, \Delta, \delta_{sJ}) \Bigg] G(\tilde{\mu}_s | z_i, \lambda, \sigma_\nu^2) & \\ \text{where, } C = \{\text{u}, \text{s}, 1, \cdots, J\} & \end{aligned}$$

In this version of the likelihood I have replaced the probabilities of the observed choices which are rooted in the theoretical model and require solving the individuals optimization problem, with a simpler function, Ω, that can be viewed as a reduced form representation of the dynamic choice probabilities. Rather than maximizing the structural parameters that best fit the data, Ω_c is a flexible specification of the states, such that,

$$\Omega_c(x_{it}, \tilde{\mu}_s) = \Pr(d_{it} = c | x_{it}, b_{it}, \tilde{\mu}_s, \Lambda, \Theta) \qquad (18)$$
$$\text{for } c \in \{\text{u}, \text{s}, 1, 2, \cdots, J\}$$

Ω_c and it's parameters have no real meaning other than they characterize the observed choices extremely well. This can be achieved with a flexible enough specification of the state variables x_{it} and $\tilde{\mu}_s$. It is important to notice that the beliefs nor the wage parameters enter into the function Ω. The reason beliefs are not included in this expression is that beliefs are simply complicated (unknown) functions of the observed data (i.e. past wages) and unobserved schooling ability. Including enough high order interactions of these variables, we will be able to distinguish how these terms influence choices, without having to solve for the beliefs explicitly. Ideally, if the choice probabilities are expressed as a robust function of the states, then including the beliefs will be redundant. The same reasoning holds for excluding the wage parameters. Since expected wages are just unknown functions of x_{it}, it's relevance in the choice probability should be accounted for with a flexible enough specification of x_{it} in Ω.

The likelihood in equation (17) avoids the previously outlined problem of having to solve the dynamic optimization problem at every guess of the parameters. By including Ω, this acts as a dynamic selection correction term and allows us to consistently estimate the other parameters in the model. However, given the other challenges outlined (high dimensional integration, joint maximization of many parameters), maximizing equation (17) directly is impractical, if not infeasible.[15]

The EM algorithm provides a tractable alternative to maximizing equation (17) directly. The benefits of the EM algorithm are twofold. First, it allows us to separately estimate the choice parameters apart from the wage parameters as well as each of the wage parameters independently. This avoids the computational challenge of maximizing all of the parameters simultaneously. Secondly, it accommodates a simple closed form solution for the integration of the occupational ability parameters such that solving for the wage parameters reduces to ordinary least squares.

The specifics of the EM algorithm are best explained by reframing the likelihood as if all of the data was observed. To do this, we define two densities pertaining to the individual's ability vector. First, let

$$g_i(\tilde{\mu}_s) = \begin{cases} 1, & \textit{if } \tilde{\mu}_s = \mu_{is} \\ 0, & \textit{otherwise} \end{cases}$$

be an individual i specific density, with a mass point of one if the input, $\tilde{\mu}_s$, equals their true schooling ability and zero otherwise. Similarly, define the individual density, f_i as,

$$f_i(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J) = \begin{cases} 1, & \textit{if } \tilde{\mu}_j = \mu_{ij} \quad \forall j \in \{1, 2, \cdots, J\} \\ 0, & \textit{otherwise} \end{cases}$$

[15]Infeasibility becomes more likely with large J because this introduces many more parameters to the model to flexibly estimate Ω.

which has a mass point of one if the input vector, $\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J$, equals their true vector of occupational abilities.

If we observed these functions, the likelihood of the data is,

$$\mathcal{L}_i = \prod_{\tilde{\mu}_s} \left[\prod_{t=1}^{T} \prod_{c \in C} \Omega_c(x_{it}, \tilde{\mu}_s)^{d_{it}=c} \prod_{\tilde{\mu}_1} \times \cdots \prod_{\tilde{\mu}_J} \left[\prod_{t=1}^{T} \prod_{j=1}^{J} \Pr(w_{it}|x_{it}, \tilde{\mu}_j, \Theta)^{d_{it}=j} \right]^{f_i(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J)} \right]^{g_i(\tilde{\mu}_s)} \tag{19}$$

The symbols $\prod_{\tilde{\mu}_1}, \prod_{\tilde{\mu}_2}, \cdots, \prod_{\tilde{\mu}_J}, \prod_{\tilde{\mu}_s}$ represent the continuous product of the inside expression evaluated for different values of $\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J, \tilde{\mu}_s$ over the support $(-\infty, \infty)$. It is the product equivalent of the integral. As $\prod_{\tilde{\mu}_s}$ cycles through all potential values of schooling ability, the likelihood is multiplied by one if it is not evaluating the individual's data at their true type. Once it cycles to the individual's true schooling ability, it returns the associated likelihood value. The series of continuous products, $\prod_{\tilde{\mu}_1}, \prod_{\tilde{\mu}_2}, \cdots, \prod_{\tilde{\mu}_J}$, is functioning in the same capacity. In the multi-dimensional case, the continuous products are cycling over all dimensions of occupational ability. When the vector is being evaluated at a point different than the true occupational ability, then it is multiplying the likelihood by one. Once it cycles to the individual's true occupational ability vector, it returns the associated likelihood value.

The benefit of maximizing the likelihood in equation (19) is that it contains only products and no sums. This means that when we take log's of the likelihood, the log operator passes completely through to the probabilities, in which the likelihood becomes a sequence of sums. Taking the log of equation (19),

$$\begin{aligned} \ln(\mathcal{L}_i) = \int_{\tilde{\mu}_s} \left(\sum_{t=1}^{T} \sum_{c \in C} (d_{it} = c) \ln(\Omega_c(x_{it}, \tilde{\mu}_s)) \right) g_i(\tilde{\mu}_s) \\ + \sum_{t=1}^{T} \sum_{j=1}^{J} (d_{it} = j) \int_{\tilde{\mu}_s} \left(\int_{\tilde{\mu}_j} \ln(\Pr(w_{it}|x_{it}, \tilde{\mu}_j, \theta_j, \sigma_j)) \right. \\ \left. \times f_i(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J) \right) g_i(\tilde{\mu}_s) \end{aligned} \tag{20}$$

Evaluating the likelihood in equation (20) is infeasible because we do know these individual f_i and g_i densities. The EM algorithm is implemented by repeatedly maximizing the likelihood in equation (20) and at each iteration, we plug in estimates of f_i and g_i ($\hat{f}_i$ and $\hat{g}_i$), which are formed from the data and the parameter estimates of the previous iteration. Dempster et al. (1977) show that solving for the parameters in this fashion is equivalent to maximizing equation (17) directly.

Given the m^{th} iterations estimates of the parameters Ω^m, Θ^m, Δ^m, δ_{sJ}^m, λ^m, and σ_ν^m, the densities $\hat{f}_i^m(\cdot)$ and $\hat{g}_i^m(\cdot)$ describe the probability distribution of the individuals unobserved abilities, $\mu_{i1}, \mu_{i2}, \cdots, \mu_{iJ}, \mu_{is}$, conditional on the observed data and the parameters. These densities are calculated with Bayes' rule, using the population distribution as the prior and updating it with the observed data.

First, the function $\hat{g}_i^m(\tilde{\mu}_s)$ is an estimated density describing the probability that an individual's schooling ability takes a particular value $\tilde{\mu}_s$. This density is formed by using the prior distribution, given by $\tilde{\mu}_s \sim \mathcal{N}(z_i\lambda, (\sigma_\nu)^2)$, and updating it with the joint probability of the observed data conditional on $\mu_{is} = \tilde{\mu}_s$. This posterior distribution will likely not have an analytical solution. In light of this, I use a discretized version of this distribution with K points of support. With K sufficiently large, this discrete distribution approaches the continuous distribution.[16]

The discrete distribution of $\hat{g}_i^m(\tilde{\mu}_s)$ is given by K discrete points, denoted by $\bar{\mu}_{sk}$ for $k \in \{1, 2, \cdots, K\}$, and K probabilities, denoted by q_{ik}^m for $k \in \{1, 2, \cdots, K\}$. The K discrete points are equidistant, covering $[-3, 3]$ standard deviations of the schooling ability distribution. The bounds associated with each point are,

$$(\underline{b}_k, \bar{b}_k) = \begin{cases} \left(-\infty, -3 + \dfrac{6}{2(K-1)}\right) & \text{if } k = 1 \\ \left(-3 + \dfrac{6(2k-3)}{2(K-1)}, -3 + \dfrac{6(2k-1)}{2(K-1)}\right) & \text{if } 1 < k < K \\ \left(3 - \dfrac{6}{2(K-1)}, +\infty\right) & \text{if } k = K \end{cases}$$

The probabilities, q_{ik}^m, represent $\Pr(\mu_{is} = \bar{\mu}_{sk})$ conditional on the m^{th} iterations estimates of the parameters. Without observing any choices or wages, the prior belief of this probability conditional

[16] In the empirical model, $K = 25$.

on the set of observables z_i is,

$$\begin{aligned}\pi_k^m(z_i, \lambda^m, \sigma_\nu^m) =& \Pr\left(\underline{b}_k < \mu_{is} < \bar{b}_k \middle| z_i, \lambda^m, \sigma_\nu^m\right) \\ =& \Phi\left(\frac{\bar{b}_k - z_i\lambda^m}{\sigma_\nu^m}\right) - \Phi\left(\frac{\underline{b}_k - z_i\lambda^m}{\sigma_\nu^m}\right)\end{aligned} \tag{21}$$

where $\Phi(\cdot)$ indicates the cumulative distribution function of the standard normal.

Letting $\Pr(\mathbf{d}_i|\cdot)$ denote the probability of all of the observed choice $\mathbf{d}_i \in \{ d_{i1}, d_{i2}, \cdots, d_{iT}\}$ and $\Pr(\mathbf{w}_i|\cdot)$ denote the probability of all of the observed wages, then using this information and the prior above, q_{ik}^m is the posterior distribution characterized by,

$$q_{ik}^m \propto \pi_k^m(z_i, \lambda^m, \sigma_\nu^m) \Pr(\mathbf{d}_i|\mathbf{x}_i, \Omega_c^m, \bar{\mu}_{sk}) \Pr(\mathbf{w}_i|\mathbf{x}_i, \Theta^m, \Delta^m, \delta_{sJ}^m, \bar{\mu}_{sk}) \tag{22}$$

where $\mathbf{x}_i \in \{x_{i1}, x_{i2}, \cdots, x_{iT}\}$ is the observed data and $q_{ik}^m = \dfrac{q_{ik}^m}{\sum_{k'=1}^K q_{ik'}^m}$.

More specifically, conditional on $\mu_{is} = \bar{\mu}_{sk}$ and the m^{th} iterations estimates of the reduced form choice probabilities, Ω_c^m, the joint probability of the observed choices in equation (22) is,

$$\Pr(\mathbf{d}_i|\mathbf{x}_i, \Omega_c^m, \bar{\mu}_{sk}) = \prod_{t=1}^T \prod_{c \in C} \Omega_c^m(x_{it}, \bar{\mu}_{sk})^{d_{it}=c} \tag{23}$$

Conditional on $\mu_{is} = \bar{\mu}_{sk}$, the probability of the observed wages, $\Pr(\mathbf{w}_i|\cdot)$, is found by expressing the distribution of $\mathbf{w}_i$ as a multi-variate normal and evaluating the pdf at the observed values. The distribution will be dictated by the individual specific career paths, so it will be unique to each worker. To give an example as to how this is constructed, assume a worker only works in periods t,t', and t'', such that $d_{it} = j$, $d_{it'} = j'$, and $d_{it''} = j''$. The joint distribution of wages for this worker's observed wages, $\mathbf{w}_i = [w_{it}, w_{it'}, w_{it''}]'$ and choices is,

$$\mathbf{w}_i \sim \mathcal{N}(\bar{\mathbf{w}}_i, \bar{\bar{\mathbf{w}}}_i) \tag{24}$$

$$\bar{\mathbf{w}}_i = \begin{bmatrix} h_{it}\theta_j^m + \bar{\mu}_{sk}\delta_{sJ_j}^m \\ h_{it'}\theta_{j'}^m + \bar{\mu}_{sk}\delta_{sJ'_j}^m \\ h_{it''}\theta_{j''}^m + \bar{\mu}_{sk}\delta_{sJ''_j}^m \end{bmatrix}$$

$$\bar{\bar{\mathbf{w}}}_i = \begin{bmatrix} \Delta_{j,j}^m - (\delta_{sJ_j}^m)^2 + (\sigma_j^m)^2 \; ; \; \Delta_{j,j'}^m - \delta_{sJ_j}^m \delta_{sJ_{j'}}^m \; ; \; \Delta_{j,j''}^m - \delta_{sJ_j}^m \delta_{sJ_{j''}}^m \\ \Delta_{j',j}^m - \delta_{sJ_{j'}}^m \delta_{sJ_j}^m \; ; \; \Delta_{j',j'}^m - (\delta_{sJ_{j'}}^m)^2 + (\sigma_{j'}^m)^2 \; ; \; \Delta_{j',j''}^m - \delta_{sJ_{j'}}^m \delta_{sJ_{j''}}^m \\ \Delta_{j'',j}^m - \delta_{sJ_{j''}}^m \delta_{sJ_j}^m \; ; \; \Delta_{j'',j'}^m - \delta_{sJ_{j''}}^m \delta_{sJ_{j'}}^m \; ; \; \Delta_{j'',j''}^m - (\delta_{sJ_{j''}}^m)^2 + (\sigma_{j''}^m)^2 \end{bmatrix}$$

The probability of the observed wages $\mathbf{w}_i$ is then found by evaluating the pdf at this point,

$$\begin{aligned} &\Pr(\mathbf{w}_i | \mathbf{x}_i, \Theta^m, \Delta^m, \delta_{sJ}^m, \bar{\mu}_{sk}) \\ &\quad = (2\pi)^{-\frac{\#\{\mathbf{w}_i\}}{2}} \det(\bar{\bar{\mathbf{w}}}_i)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{w}_i - \bar{\mathbf{w}}_i)' \bar{\bar{\mathbf{w}}}_i^{-1} (\mathbf{w}_i - \bar{\mathbf{w}}_i)\right) \end{aligned} \tag{25}$$

where $\#\{\mathbf{w}_i\}$ is the number of wage observations.

Turing now to the other density, $\hat{f}_i^m$, which represents the distribution of the individual's occupational ability. If we condition on the observed data and the m^{th} iterations estimate of the parameters,

$$\begin{aligned} &\hat{f}_i^m(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J | \tilde{\mu}_s) \\ &\quad = \Pr(\mu_{ij} = \tilde{\mu}_j \; \forall j \in \{1, 2, \cdots, J\} | \mathbf{w}_i, \mathbf{x}_i, \tilde{\mu}_s, \Theta^m, \Delta^m, \delta_{sJ}^m) \end{aligned} \tag{26}$$

What is important about the density in equation (26) is first the schooling ability, $\tilde{\mu}_s$, is treated as data, so the density is conditional on this value. The second feature is that the choices are not included when we condition on the data (past wages). This goes back to the fact that individuals condition their choices on their beliefs and not their real values of ability. Since beliefs are simply a function of the past wages and their schooling ability, once we condition on these values in equation (26), choices add no new information about the real value of abilities. Finally, the expression for the density function, $\hat{f}_i^m$, has a very familiar form. Since the updates in the EM algorithm are formed using Bayes' rule, this distribution will be the same distribution as the beliefs the individuals have in the model once they condition on all of their wages (since the individuals are Bayesian updaters also).

For each potential value of schooling $\bar{\mu}_{sk}$ the density $\hat{f}^m_{ik}$ is defined as,

$$\hat{f}^m_{ik}(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J) \sim \mathcal{N}\big(\hat{\mathbf{E}}^m_{ik}, \hat{\mathbf{V}}^m_{ik}\big) \tag{27}$$

where,

$$\hat{\mathbf{E}}^m_{ik} = \Big(\big(\Delta^m - \delta^m_{sJ}\delta^{m'}_{sJ}\big)^{-1} + \Psi_i\Big)^{-1}\Big(\big(\Delta^m - \delta^m_{sJ}\delta^{m'}_{sJ}\big)^{-1}\big(\delta^m_{sJ}\bar{\mu}_{sk}\big) + \mathbf{W}_i\Big)$$

$$\hat{\mathbf{V}}^m_i(\tilde{\mu}_s) = \Big(\big(\Delta^m - \delta^m_{sJ}\delta^{m'}_{sJ}\big)^{-1} + \Psi_i\Big)^{-1}$$

$$\Psi_{iJ\text{x}J} = \begin{cases} \Psi_{i_{j,j}} = \dfrac{\sum_{t=1}^T (d_{it} = j)}{(\sigma^m_j)^2} & , \quad \forall j \in \{1, 2, \cdots, J\} \\ 0 & , \quad \textit{everywhere else} \end{cases}$$

$$\mathbf{W}_{iJ\text{x}1} = \Big\{ \ \mathbf{W}_{i_j} = \frac{\sum_{t=1}^T (d_{it} = j)(w_{it} - h_{it}\theta^m_j)}{(\sigma^m_j)^2} \quad , \quad \forall j \in \{1, 2, \cdots, J\}$$

$\hat{\mathbf{E}}^m_{ik}$ is the estimated mean of the occupational ability given: schooling ability equals $\bar{\mu}_{sk}$, the parameters at iteration m, and the observed data. $\hat{\mathbf{V}}^m_{ik}$ is the estimated covariance. The principal of the EM algorithm is that as we collect more data on an individual, the estimated covariance goes to zero and the estimated mean goes to their true value.

$\hat{f}^m_{ik}$ and q^m_{ik} are individual densities. We can aggregate these up to get the population densities, Δ and δ_{sJ}. The estimates at the $(m+1)$ iteration for these distribution parameters are,

$$\Delta^{m+1}_{(j,j')} = \frac{1}{N}\sum_{i=1}^N \sum_{k=1}^K q^m_{ik}\left[\hat{\mathbf{V}}^m_{ik_{(j,j')}} + \hat{\mathbf{E}}^m_{ik_{(j)}}\hat{\mathbf{E}}^m_{ik_{(j')}}\right] \tag{28}$$

$$\delta^{m+1}_{sJ_j} = \frac{1}{N}\sum_{i=1}^N \sum_{k=1}^K q^m_{ik}\left[\hat{\mathbf{E}}^m_{ik_{(j)}}\bar{\mu}_{sk}\right]$$

where $\hat{\mathbf{V}}^m_{ik_{(j,j')}}$ represents the j^{th} row and j'^{th} column of the covariance matrix $\hat{\mathbf{V}}^m_{ik}$, and $\hat{\mathbf{E}}^m_{ik_{(j)}}$ refers to the j^{th} element of the vector $\hat{\mathbf{E}}^m_{ik}$.

Finally the parameters for the initial conditions equation is updated using q_{ik}^m as,

$$\lambda^{m+1} = \underset{\lambda,\sigma_\nu}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}^m \ln\left(\Phi\left(\frac{\overline{\mathbf{b}}_k - z_i\lambda}{\sigma_\nu}\right) - \Phi\left(\frac{\underline{\mathbf{b}}_k - z_i\lambda}{\sigma_\nu}\right)\right) \tag{29}$$

$$\begin{aligned} \text{s.t.} \quad & \mathbb{E}(z_i\lambda) = 0 \\ & \mathbb{E}((z_i\lambda)^2) + \sigma_\nu^2 = 1 \\ & \Phi \text{ is the cdf for the standard normal} \end{aligned}$$

The estimator for the initial conditions is an interval estimator, reflecting the fact that we have a continuous latent variable μ_{is}, but we only know the probability that is falls within certain bounds (i.e. q_{ik} represents the probability that $\mu_{is} \in [\underline{\mathbf{b}}_k, \overline{\mathbf{b}}_k]$). The constrains enforce that the necessary normalizations bind.

Implementing the EM algorithm requires the repeated computation of two steps, the expectation (E-step) and the maximization (M-step). So far I have completely characterized the expectation step, which describes the updates for $\hat{f}_{ik}^m$, q_{ik}^m, Δ^m, δ_{sJ}^m, λ^m, and σ_ν^m. These values are calculated outside of the maximization routine following the formulas outlined above. The remaining parameters in the model are the functions for the reduced form choice probabilities, Ω_c, and the parameters of the wage equation, Θ. To update these parameters, we maximize the likelihood in equation (20), plugging in the values derived in the expectation step. Equation (20) still has the continuous integration of schooling ability, so I replace it with the discrete approximation, so that the new log-likelihood is,

$$\begin{aligned} \ln(\mathcal{L}_i) = & \sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T}\sum_{c\in C}(d_{it}=c)\ln(\Omega_c(x_{it},\bar{\mu}_{sk})) \\ & + \sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T}\sum_{j=1}^{J}(d_{it}=j)\bigg(\int_{\tilde{\mu}_j} \ln(\Pr(w_{it}|x_{it},\tilde{\mu}_j,\theta_j,\sigma_j)) \\ & \quad \hat{f}_{ik}^m(\tilde{\mu}_1,\tilde{\mu}_2,\cdots,\tilde{\mu}_J)\bigg) \end{aligned} \tag{30}$$

Two features of this likelihood function make it appealing to maximize. First the parameters Ω^c, and the wage parameters, Θ, are additively separable. This means that we can maximize these two components of the likelihood separately, which avoids the computational burden of jointly maximizing all of the parameters. The update of the reduced form choice probabilities Ω_c^{m+1}

solve[17],

$$\Omega_c^{m+1} = \underset{\Omega_c}{\text{argmax}} \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T}\sum_{c\in C}(d_{it}=c)\ln(\Omega_c(x_{it},\bar{\mu}_{sk})) \tag{31}$$

The second attractive feature of the likelihood in equation (30) is that the wage observations are no longer collectively tied together through a multivariate integral. Instead each wage is only integrated over the occupational ability in that single occupation. We can replace the density, $\hat{f}_{ik}^m(\tilde{\mu}_1,\tilde{\mu}_2,\cdots,\tilde{\mu}_J)$ with it's relevant marginal density, $\hat{f}_{ik}^m(\tilde{\mu}_j) \sim \mathcal{N}(\hat{\mathbf{E}}_{ik_{(j)}}^m,\ \hat{\mathbf{V}}_{ik_{(j,j)}}^m)$, where the expressions $\hat{\mathbf{E}}_{ik_{(j)}}^m$ and $\hat{\mathbf{V}}_{ik_{(j,j)}}^m$ are defined in equation (28). Writing the probability of the observed wage with respect to the unobserved technology shock, $\eta \sim \mathcal{N}(0,\sigma_j^2)$ as,

$$\Pr(w_{it}|x_{it},\tilde{\mu}_j,\theta_j,\sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{(w_{it}-h_{it}\theta_j-\tilde{\mu}_j)^2}{2\sigma_j^2}\right) \tag{32}$$

and taking the log of this probability and integrating over the unobserved occupational ability, yields,

$$\begin{aligned}
&\int_{\tilde{\mu}_j} \ln(\Pr(w_{it}|x_{it},\tilde{\mu}_j,\theta_j,\sigma_j))\hat{f}_{ik}^m(\tilde{\mu}_j) = \\
&= \int_{\tilde{\mu}_j} \ln\left(\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left(-\frac{(w_{it}-h_{it}\theta_j-\tilde{\mu}_j)^2}{2\sigma_j^2}\right)\right)\hat{f}_{ik}^m(\tilde{\mu}_j) \\
&= \int_{\tilde{\mu}_j}\left[-\frac{1}{2}\ln(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}\left(w_{it}-h_{it}\theta_j-\tilde{\mu}_j\right)^2\right]\hat{f}_{ik}^m(\tilde{\mu}_j) \\
&= -\frac{1}{2}\ln(2\pi\sigma_j^2) - \frac{1}{2\sigma_j^2}\left(\hat{\mathbf{V}}_{ik_{(j,j')}}^m + \left(w_{it}-h_{it}\theta_j-\hat{\mathbf{E}}_{ik_{(j)}}^m\right)^2\right)
\end{aligned} \tag{33}$$

Equation (33) has a closed form solution for the integration of the unobserved occupational ability. We can take advantage again of the additive separability in equation (30) and estimate the wage parameters separately for each occupation. Plugging equation (33) into the the likelihood in equation (30) and taking derivatives with respect to θ_j and setting it equal to zero, gives the

[17] For the empirical work, the function Ω_c is flexible a logit of the sate variables, x_{it} and $\bar{\mu}_{sk}$.

objective function for the updates of θ_j^{m+1} as,

$$\theta_j^{m+1} = \underset{\theta_j}{\text{argmax}} \sum_{i=1}^{N} \sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T} (d_{it} = j) \left[w_{it} - h_{it}\theta_j - \hat{\mathbf{E}}_{ik_{(j)}}^m \right]^2 \tag{34}$$

which is simply solved using weighted ordinary least squares.

Equation (34) provides a clear illustration of how the estimator addresses selection. $\hat{\mathbf{E}}_{ik_{(j)}}^m$ is an individual specific mean. If individuals with high ability are more likely to have high tenure, then this effect will be corrected for by the mean, rather than absorbed into the returns to tenure.

Taking the derivative with respect to σ_j gives a closed form expression for updating σ_j^{m+1}.

$$(\sigma_j^{m+1})^2 = \tag{35}$$
$$\frac{\sum_{i=1}^{N} \sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T} (d_{it} = j) \left(\hat{\mathbf{V}}_{ik_{(j,j')}}^m + \left(w_{it} - h_{it}\theta_j^{m+1} - \hat{\mathbf{E}}_{ik_{(j)}}^m \right)^2 \right)}{\sum_{i=1}^{N} \sum_{k=1}^{K} q_{ik}^m \sum_{t=1}^{T} (d_{it} = j)}$$

Summarizing the Algorithm

The EM algorithm is used in the first stage to solve for all of the model parameters except the structural utility parameters. Instead, a reduced form logit model is used to account for dynamic selection, postponing the estimation of the structural model to the second stage. The EM algorithm breaks the joint likelihood in the parameters, allowing us to solve for the choice parameters and the wage parameters separately. In addition, we are able to eliminate the multiple integration of the occupational abilities and substituted it with a closed form for the expectation so that the wage parameters are easily maximized using weighted ordinary least squares.

Step 1. Given the m^{th} iteration parameter estimates, Ω_c^m, Θ^m, Δ^m, δ_{sJ}^m, λ^m, and σ_ν^m solve for the probability that individual i's true schooling type is $\bar{\mu}_{sk}$, q_{ik}^m for each $k \in \{1, 2, \cdots, K\}$.

Step 2. Solve for $\hat{f}_{ik}^m(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J)$, the individual's probability density for their occupational ability conditional on their schooling ability being $\bar{\mu}_{sk}$.

Step 3. Using q_{ik}^m and the wage parameters, Θ^m, solve for the beliefs over occupational ability b_{it}^m for each $\bar{\mu}_{sk}$.

Step 4. Update the population parameters Δ^{m+1}, δ_{sJ}^{m+1}, λ^{m+1}, and σ_{ν}^{m+1} with $\hat{f}_{ik}^{m}(\tilde{\mu}_1, \tilde{\mu}_2, \cdots, \tilde{\mu}_J)$ and q_{ik}^{m}.

Step 5. Maximize the reduced form probabilities Ω_c^{m+1} given q_{ik}^{m}, the beliefs b_{it}^{m}, and expected wages using Θ^m. (Although it is sufficient to flexibly express the reduced form choice probabilities as only a function of the data, since the EM allows us to get estimates for the unobserved beliefs and expected wages, these will be included also).

Step 6. Do weighted OLS for each occupation $j \in \{1, 2, \cdots, J\}$ to update θ_j^{m+1} and σ_j^{m+1}

The algorithm repeats steps 1-6 until the change in the log-likelihood is less than some convergence criteria $\left(\sum_{i=1}^{N} \ln(\mathcal{L}_i)\right)^{m+1} - \left(\sum_{i=1}^{N} \ln(\mathcal{L}_i)\right)^{m} < 1e-4$.

3.3 Second Stage: Solving the Dynamic Discrete Choice Problem

With the first stage estiamtes, we now move to estimating the structural utility parameters. From the first stage, we have recovered consistent estimates of the beliefs, $\hat{b}_{it}$ and the wage parameters, $\hat{\Theta}$: all which play an important role in solving the structural model. Although the problem is simplified by recovering these parameters outside of the structural model, estimating the dynamic discrete choice model is still extremely complicated because evaluating the choice probabilities requires solving the individual's expected future value associated with each choice. In order to address this overwhelming state space problem, I utilize the method of conditional choice probabilities (CCP's) developed by Hotz and Miller (1993) and Arcidiacono and Miller (2010) to avoid having to solve the model at every point in the state space.

To solve for the structural parameters, we need to maximize the log-likelihood,

$$\ln(\mathcal{L}_i) = \sum_{k=1}^{K} \hat{q}_{ik} \sum_{t=1}^{T} \ln(\Pr(d_{it}|x_{it}, \hat{b}_{it}, \bar{\mu}_{sk}, \Lambda, \hat{\Theta})) \tag{36}$$

where Λ are the structural utility parameters and the hatted variables are estimated in the first stage. The choice probabilities in equation (36) are taken with respect to the unobserved, per-period random utility shock ε. Assuming these unobservables are distributed *i.i.d.* type-I extreme value, we can use the conditional value functions in equation (10) in section 2.3 to express these probabilities.

$$\Pr(d_{it} = c | x_{it}, \hat{b}_{it}, \bar{\mu}_{sk}, \Lambda, \hat{\Theta}) = \frac{\exp(v_{ict}(\cdot))}{\sum_{c'} \exp(v_{ic't}(\cdot))} \tag{37}$$

where

$$v_{ict}(\cdot) = u_{ict}(\mathcal{S}_{it}) + \beta \mathbb{E}[V_{t+1}(\mathcal{S}_{it+1}, \boldsymbol{\varepsilon}_{t+1}) | \mathcal{S}_{it}, d_{it} = c]$$

The expected future value term $\mathbb{E}[V_{t+1}(\cdot)]$ is a complicated function of all potential future career sequences, where the expectation is taken over all possible ability signals in the current and future periods, as well as next periods and all future non-pecuniary choice specific utility shock, ε. Rust (1987) shows that given that these utility shocks are additively separable in the utility function and that they are type-I extreme value, the expectation with respect to next periods utility shock has a closed form expression as the log sum of the exponential of next periods conditional value functions,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} V_{t+1}(\mathcal{S}, \boldsymbol{\varepsilon}) = \ln \left(\sum_{c'} \exp(v_{ic't+1}(\mathcal{S})) \right) + \iota,$$

where ι is Euler's constant

The term $v_{ic't+1}$ represents the conditional value function of choice c' in the next period and the sum is over all feasible choices.

Hotz and Miller (1993) show that if the expression inside of the log is multiplied top and bottom by $\exp(v_{ict+1}(\mathcal{S}))$, for any choice c, then the expected future value term can be equivalently expressed as the log of the conditional choice probability, plus the choice specific value function.

$$\begin{aligned}
\mathbb{E}_{\varepsilon} V_{t+1}(\mathcal{S}, \varepsilon) &= \ln\left(\frac{\left(\sum_{c'} \exp(v_{ic't+1}(\mathcal{S}))\right) \exp(v_{ict+1}(\mathcal{S}))}{\exp(v_{ict+1}(\mathcal{S}))}\right) + \iota \\
&= \ln\left(\frac{\sum_{c'} \exp(v_{ic't+1}(\mathcal{S}))}{\exp(v_{ict+1}(\mathcal{S}))}\right) + v_{ict+1}(\mathcal{S}) + \iota \\
&= \ln\left(\Omega_c(\mathcal{S})^{-1}\right) + v_{ict+1}(\mathcal{S}) + \iota \\
&= -\ln\left(\Omega_c(\mathcal{S})\right) + v_{ict+1}(\mathcal{S}) + \iota \qquad (38)
\end{aligned}$$

where $\Omega_c(\mathcal{S})$ is the conditional choice probability of making choice c given state $\mathcal{S}$.

While the Hotz and Miller (1993) choice probability inversion is helpful in solving for $\mathbb{E}[V_{t+1}(\cdot)]$, we are left to evaluate the next periods choice specific valuation function, $v_{ict+1}(\mathcal{S})$, which contains the expectation of the two period ahead value function $\mathbb{E}V_{it+2}(\mathcal{S}', \varepsilon' | \mathcal{S}, d_{it+1} = c)$. Again, not being able to compute this value, we need an alternative representation. One approach would be to apply equation (38) again, however it is clear that using CCP's in this manner will require us taking the problem out all the way to the terminal period T.

Arcidiacono and Miller (2010) demonstrate that dynamic models that possess the property of finite dependence can utilize CCP's in a way which does not require solving the model to the terminal period. In fact, with finite dependence, only a few period ahead CCP's are required. Finite dependence takes advantage of two features of the problem. First, equation (38) holds true for any choice c, meaning that with $J + 2$ choices, there are $J + 2$ equivalent ways of expressing $\mathbb{E}[V_{t+1}(\cdot)]$. The second feature is that in estimating any discrete choice model, only differences in utilities matter. The idea proposed by Arcidiacono and Miller (2010) is that if we strategically express the expected future value term as a sequence of choices such that when we difference two value functions the remaining expected future value differences out, then we can estimate the structural model, requiring only a few period ahead CCP's. In the case of the model proposed here, we need to look at two period ahead conditional choice probabilities.

In estimation, I normalize all of the value functions against the unemployment choice. Therefore the probability functions are

$$\Pr(d_{it} = c | x_{it}, \hat{b}_{it}, \bar{\mu}_{sk}, \Lambda, \hat{\Theta}) = \frac{\exp(v_{ict}(\cdot)) - \exp(v_{i\mathrm{u}t}(\cdot))}{\sum_{c'}(\exp(v_{ic't}(\cdot)) - \exp(v_{i\mathrm{u}t}(\cdot)))} \tag{39}$$

Finite dependence suggests expressing the above conditional value functions strategically so that the three period ahead expected future value term differences out. This strategic differencing is done for every possible choice. To demonstrate how this works, I will outline the expression for the conditional value of working in occupation 1, (i.e. $c = 1$), subtracting the conditional value function of unemployment in the same period. Given there are $J+2$ choices which can be made in the next period, we can express today's value function of choosing occupation 1 in $J+2$ ways. If we apply the conditional choice probability inversion to replace the expected future value in the next period, then there are $(J+2)^2$ equivalent ways we can express the conditional value of choosing occupation 1 today. One of the $(J+2)^2$ equivalent ways is expressed as,

$$\begin{aligned} v_{i1t}(\mathcal{S}^0) = u_{i1t}(\mathcal{S}^0) + \beta \int_{\zeta_1} \Big[& - \ln(\Omega_{\mathrm{u}}(\mathcal{S}^{a1})) + u_{i\mathrm{u}t+1}(\mathcal{S}^{a1}) \\ & + \beta \left(- \ln(\Omega_{\mathrm{u}}(\mathcal{S}^{a2})) + u_{i\mathrm{u}t+2}(\mathcal{S}^{a2}) + \beta \mathbb{E} V_{it+3}(\mathcal{S}^{a3}) \right) \Big] \end{aligned} \tag{40}$$

where,

$$\begin{aligned} \mathcal{S}^{a1} &= \{\mathcal{S}^0, (age_{it}+1), (expr_j+1), d'_{it} = 1, \zeta_1\} \\ \mathcal{S}^{a2} &= \{\mathcal{S}^0, (age_{it}+2), (expr_j+1), d'_{it+1} = \mathrm{u}, \zeta_1\} \\ \mathcal{S}^{a3} &= \{\mathcal{S}^0, (age_{it}+3), (expr_j+1), d'_{it+2} = \mathrm{u}, \zeta_1\} \\ \zeta_1 &\sim \mathcal{N}(\mathbf{E}_{it}(\mu_{i1}), \mathbf{V}_{it}(\mu_{i1}) + \sigma_1^2) \end{aligned}$$

which represents the sequence $\{d'_{it} = \text{occupation } 1, d'_{it+1} = \text{unemployment}, d'_{it+2} = \text{unemployment}\}$, where the primes ($'$) represent a potential sequence. $\mathcal{S}^{a1}, \mathcal{S}^{a2}, \mathcal{S}^{a3}$ represent how the state space evolves one, two, and three periods ahead respectively, given this potential sequence. For the worker, what is relevant three periods away, is that they will be three years older than they were at $\mathcal{S}^0$, they will have one more year of occupation 1 experience, they will have been unemployed in the previous period, and they will have received one ability signal for occupation 1.

Likewise, there are $(J+2)^2$ equivalent ways we can define the conditional value of being unemployed in the current period. One such way is the potential path, $\{d'_{it} = \text{unemployment}, d'_{it+1} =$

occupation $1, d'_{it+2}=$ unemployment$\}$, which is,

$$v_{i\mathrm{u}t}(\mathcal{S}^0) = u_{i\mathrm{u}t}(\mathcal{S}^0) + \beta \int_{\zeta_1} \Big[-\ln(\Omega_1(\mathcal{S}^{b1})) + u_{i1t+1}(\mathcal{S}^{b1}) \\ + \beta\Big(-\ln(\Omega_{\mathrm{u}}(\mathcal{S}^{b2})) + u_{i\mathrm{u}t+2}(\mathcal{S}^{b2}) + \beta \mathbb{E} V_{it+3}(\mathcal{S}^{b3}) \Big) \Big] \tag{41}$$

where,

$$\mathcal{S}^{b1} = \{\mathcal{S}^0, (age_{it}+1), (expr_j+0), d'_{it}=\mathrm{u}\}$$

$$\mathcal{S}^{b2} = \{\mathcal{S}^0, (age_{it}+2), (expr_j+1), d'_{it+1}=1, \zeta_1\}$$

$$\mathcal{S}^{b3} = \{\mathcal{S}^0, (age_{it}+3), (expr_j+1), d'_{it+2}=\mathrm{u}, \zeta_1\}$$

$$\zeta_1 \sim \mathcal{N}(\mathbf{E}_{it}(\mu_{i1}), \mathbf{V}_{it}(\mu_{i1}) + \sigma_1^2)$$

Finite dependence shows up in the fact that the expected state vector three periods away for both sequences are equivalent (i.e. $\beta^3\int_{\zeta_1}\mathbb{E}V_{it+3}(\mathcal{S}^{a3}) = \beta^3\int_{\zeta_1}\mathbb{E}V_{it+3}(\mathcal{S}^{b3})$). When we difference these two expressions of the conditional value functions to estimate the parameters, the unknown value will difference out. This leaves for estimation,

$$v_{i1t}(S^0) - v_{i\mathrm{u}t}(S^0) = u_{i1t}(\mathcal{S}^0) - \beta\int_{\zeta_1}\ln(\Omega_{\mathrm{u}}(\mathcal{S}^{a1})) - \beta^2\int_{\zeta_1}\ln(\Omega_{\mathrm{u}}(\mathcal{S}^{a2})) \\ - \Big[-\beta\int_{\zeta_1}\ln(\Omega_1(\mathcal{S}^{b1})) + \beta\int_{\zeta_1} u_{i1t+1}(\mathcal{S}^{b1}) - \beta^2\int_{\zeta_1}\ln(\Omega_{\mathrm{u}}(\mathcal{S}^{b2})) \Big] \tag{42}$$

where the the flow utility for unemployment is normalized to zero (i.e. $u_{i\mathrm{u}t}=0$).

This representation of the differenced value function reduces down to an expression which contains only the contemporaneous utility functions and conditional choice probabilities. Repeating this strategic differencing for each of the other choices, we can get an expression for the structural utility model without directly solving the optimization problem, or doing any backwards recursion from the terminal period.

The conditional choice probabilities were estimated in the first stage. Plugging these values in as data, I use the expressions above for the differenced conditional value functions to solve for the structural utility parameters,

$$\hat{\Lambda} = \underset{\Lambda}{\operatorname{argmax}} \sum_{k=1}^{K} \hat{q}_{ik} \sum_{t=1}^{T} \ln(\Pr(d_{it}|x_{it}, \hat{b}_{it}, \bar{\mu}_{sk}, \Lambda, \hat{\Theta}, \hat{\Omega}_c)) \tag{43}$$

4 Data Extract and Summary: NLSY97

The model is estimated using the 1997 cohort of the National Longitudinal Survey of Youth (NLSY97). The NLSY97 is a nationally representative sample of men and women who were born between 1980 and 1984. These individuals were between the ages of 12 and 18 at the time of their first interview in 1997, and respondents are interviewed annually. I use rounds 1 to round 12 for the analysis, where the respondents are between the ages of 23 and 29 in round 12.[18] On an annual basis, interviewers collect detailed information on the respondent's education and employment activities retrospectively to the date of the last interview.

Based on the responses to the retrospective education and employment questions, each year an individual is assigned to one of twelve mutually exclusive activities: unemployment, schooling, or employment in one of 10 occupations { 1-Management, business, and financial operations occupations, 2-Professional and related occupations, 3-Other Service occupations, 4-Food Service occupations, 5-Sales and related occupations, 6-Office and administrative support occupations, 7-Construction and extraction occupations, 8-Installation, maintenance, and repair occupations, 9-Production occupations, 10-Transportation and material moving occupations}.

The assignment of activities was done sequentially, starting with education. If the individual reports attending school for five out of the 12 months in a year for high schoolers (four out of 12 months for college students) and reported an increase in their highest grade completed between interviews, then they were categorized as enrolled in school for the period.[19] If information is not available for the entire year, for example in their last interview round, then the attendance in the available months was converted to a 12 month equivalent and the same criteria applied. In addition, in the last round, no information is available regarding whether they completed a grade that year,

[18] Round 13 is the most current round available. This survey was conducted in 2009. Including year 2009 had drastic effects on the unemployment rate of these workers, which calls into question the stationarity assumption of the structural choice parameters. Conceivably it is possible to allow for aggregate shocks into the model, but that is outside the scope of this paper. However, the round 13 information was used to supplement the round 12 data where possible to provide a more complete picture of the individuals activities in round 12, rather than relying on a partial years worth of data.

[19] In a few cases grade completion was hard coded to more accurately reflect school progression. For example, if the individual reported attending college for two years but did not report a grade increase in the subsequent year, but reported skipping a grade two periods away (i.e. 12, 12, 14), then the middle year was re-coded as 13, assuming they were continually in school rather than dropping out and then skipping a grade. Only a handful of records required this scrutiny.

Table 1: Occupation Categories

Occ. Cat.	2002 3-Digit Census Code	Description
1	0010-0950	**Management, business, and financial operations occupations**
		0010-0430 Management occupations
		0500-0950 Business and financial operations occupations
2	1000-3540	**Professional and related occupations**
		1000-1240 Computer and mathematical occupations
		1300-1560 Architecture and engineering occupations
		1600-1960 Life, physical, and social science occupations
		2000-2060 Community and social services occupations
		2100-2150 Legal occupations
		2200-2550 Education, training, and library occupations
		2600-2960 Arts, design, entertainment, sports, and media occupations
		3000-3540 Healthcare practitioner and technical occupations
3	3600-3950; 4200-4650	**Non-Food Service Occupations**
		3600-3650 Healthcare support occupations
		3700-3950 Protective service occupations
		4200-4250 Building and grounds cleaning and maintenance occupations
		4300-4650 Personal care and service occupations
4	4000-4160	**Food preparation and serving related occupations**
5	4700-4960	**Sales and related occupations**
6	5000-5930	**Office and administrative support occupations**
7	6000-6940	**Construction and extraction occupations**
		6000-6130 Farming, fishing, and forestry occupations
		6200-6940 Construction and extraction occupations
8	7000-7620	**Installation, maintenance, and repair occupations**
9	7700-8960	**Production occupations**
10	9000-9750	**Transportation and material moving occupations**

so attendance was the sole criteria in determining the schooling category.

If the respondent did not meet the school enrollment criteria and reported working at least 25 weeks in the year in a job working at least 20 hours per week, then the individual was considered employed. Again, if the individual was not present for the entire year, the weeks observed where converted to an annual equivalent. If the individual worked multiple full-time jobs, then the attributes of the job with the most full-time weeks was assigned for that year. For each job, the respondent identifies the 3-digit 2002 Census occupation code and hourly compensation rate. Given the reported 3-digit occupation, the individual is assigned to one of the 10 occupation categories following table 1. Wage is assigned using the hourly compensation rate of pay, which includes all forms of monetary compensation affiliated with the job and is deflated using the consumer price index to year 2002 dollars.

The main purpose of utilizing the estimation strategy outlined in the previous section is that the computational difficulty only grows linearly in the number of occupations in the model. In this sense, the empirical model can accommodate a larger choice set than the ten aggregated occupation codes described above. However, what prohibits us from analyzing occupations at a finer level of detail is sample size. In the current specification all of the wage parameters and choice parameters are occupation specific. Including the cross-occupational correlation terms, the technology shock standard error, and other parameters, each occupation is associated with about 27 parameters. Without additional assumptions (e.g. forcing the wage returns to education to be constant across occupations), we require a reasonable amount of observations in these occupations to identify the parameters. The ten occupation categories where drawn such that each occupation category provided at least 350 wage observations and categories were consistent with the 2002 occupation classification system. Table 1 shows how the 2002, 1-digit census occupation codes map to the ten occupations used in the empirical analysis. Six of the 10 occupation groups contain a single 1-digit occupation code. Two of the ten only combine two 1-digit occupation codes. Only professional and other service occupations combine multiple 1-digit occupation codes.

If an individual is not assigned to schooling or one of the 10 occupation employment activities, then they are assigned to unemployment (or home production).

The analysis focuses on the white male cross-sectional sample from the data. This group consists of 2,284 individuals. The discrete decision period corresponds to the school year (September to August). The decisions of individuals are continually tracked from age 16 (t=1) until round 12 unless any one of five events occur: 1) the gap between interview dates exceeds 16 months, 2) if their highest grade completed or degree completed is invalid from the survey data, 3) the individual's primary activity was employment, but no occupation was reported or the reported wage was less than \$5 per hour or greater than \$100 per hour, 4) they joined the military, or 5) reported completing 20 or more years of education.[20] Of the original population, 2,125 individuals survive to age 16, with the final sample containing 15,198 person year observations, averaging 7.15 years per individual.

Table 2 summarizes the distribution of activities by age, as well as the number of observations at each age. As expected, unemployment for these young workers is relatively high ranging between

[20] There appear to be a large fraction of medical students who meet the employment criteria with 20 years education. Mean wages increase monotonically across education levels except for 19 years to 20 years where the average wage falls 28%. These outliers have a strong effect on the returns to education, so these 20 observations are dropped.

Table 2: Distribution of Choices By Age

Age	Unemp.	School	Man./ Bus.	Prof.	Service	Food	Sales	Office/ Admin.	Const-ruct.	Maint./ Repair	Prod-uction	Transp.	Obs.
16	3%	93%	0%	0%	0%	1%	0%	0%	1%	0%	0%	0%	2125
17	6%	85%	0%	0%	0%	2%	1%	1%	2%	1%	1%	1%	1933
18	10%	60%	0%	1%	2%	4%	4%	2%	6%	2%	3%	5%	1768
19	12%	45%	0%	1%	3%	5%	5%	4%	9%	3%	5%	8%	1600
20	12%	41%	1%	2%	3%	6%	5%	4%	10%	3%	5%	7%	1471
21	11%	39%	1%	3%	4%	6%	5%	5%	11%	4%	5%	7%	1338
22	13%	25%	4%	6%	5%	6%	7%	6%	10%	5%	6%	7%	1240
23	12%	16%	6%	10%	4%	6%	9%	7%	11%	5%	6%	8%	1149
24	11%	12%	7%	12%	4%	5%	10%	8%	12%	6%	6%	8%	1005
25	12%	10%	8%	14%	5%	4%	11%	8%	11%	6%	5%	7%	728
26	15%	9%	10%	13%	3%	4%	9%	7%	12%	6%	7%	6%	482
27	11%	10%	13%	13%	5%	3%	9%	6%	12%	8%	4%	5%	263
28	14%	5%	18%	12%	5%	3%	14%	5%	10%	6%	5%	2%	96

10-15%. The share of employment is not even across the 10 occupations. At age 24 when about half of the population is still in the sample, the largest shares of employment are in professional, sales, or construction occupations. Management, office, and transportation have about average representation. And, service, food, maintenance, and production have below average shares.

Table 3 describes in greater detail the employment activities of white males between the ages of 16 and 28. For each of the ten occupation categories, table 3 reports the top four, 3-digit detailed occupation codes, as well as the percentage of observed employment records in those occupations. For food, sales, and transportation occupations, the top four detailed occupations represent more than 70% of the employment records in those categories, implying that white males are more concentrated in a handful of 3-digit occupation codes. Comparably, only 25% of employment records are represented by the top four detailed occupations in professional occupations, suggesting that workers are much more spread across the 3-digit detail occupations in professional occupations.

Previous papers modeling career decisions, (e.g. Neal (1999); Pavan (2009); Kambourov and Manovskii (2009)) using the NLSY79 or Panel Study of Income Dynamics (PSID) data have documented the high potential for measurement error when the reported occupation code is take directly from the data. To avoid counting false career changes these papers impose a number of edits on the occupational data. The primary edit is to not consider any occupational change unless it is accompanied by a change in employer. Yamaguchi (2010) points out that this may be an undesirable restriction on the data as it is likely to exclude important career changes as individuals are promoted within the firm.

The NLYS97 is unique from the NLSY79 and PSID in that it likely does not suffer from systematic measurement error in the reported occupation. The NLSY97 is conducted with a computer-assisted interview system which allows interviewers to reference back to the responses of their previous years interview. Interviewees are first read their previous years job description and are asked if that continues to define their job function. The occupation code only changes if they report a change in duties. Pavan (2009), using the NLSY79 data, cites evidence of spurious reported occupation changes by the fact that 40% of individuals remaining with the same employer in consecutive periods report a change in 3-digit occupations. The analogous figure for the NLSY97 data at the

Table 3: Description of Occupation Categories

Occupation Categories (# of 3- digit census codes)	Top-Four, 3-digit Occupations (% of obs. in category)	Num. of Obs.	Percent of Obs. in Top-Four
Management, business, and financial operations occupations (52)	Managers, all other (11.8) Accountants and Auditors (9.1) Construction Managers (8.9) Management Analysts (8.6)	373	38.4%
Professional and related occupations (123)	Computer software engineers (7.7) Computer Support Specialist (6.3) Secondary School teachers (5.7) Network System/Data analysts (4.8)	600	24.5%
Non-Food Service Occupations (49)	Grounds worker (21.1) Janitor (15.3) Security Guard (14.1) Nursing/Home health aides (10.2)	412	60.7%
Food preparation and serving related occupations (13)	Cooks (35.7) Waiters (16.0) Sup. Food Prep. (11.1) Food Prep. (9.0)	620	71.8%
Sales and related occupations (18)	Retail salesperson (28.0) Cashier (22.0) Sup. Retail sales (19.6) Sales rep. whls/mfg (6.2)	724	75.8%
Office and administrative support occupations (51)	Stock clerk (30.8) Customer Service rep (13.8) Shipping/Rec. clerk (10.4) Dispatcher (4.5)	559	59.4%
Construction and extraction occupations (48)	Carpenters (24.2) Construction laborer (22.0) Electrician (7.6) Plumbers/pipfitters (6.6)	1087	60.5%
Installation, maintenance, and repair occupations (36)	Auto tech/ mechanic (24.3) HVAC mechanic/installer (10.0) Computer/office machine repair (7.3) Auto body repair (6.9)	452	48.5%
Production occupations (81)	Production worker (15.0) Assembler/frabricator (12.3) Welding/soldering worker (9.5) Metal worker/ plastic worker (7.0)	547	43.7%
Transportation and material moving occupations (34)	Laborer and freight movers (32.8) Driver/Sales, truck driver (22.7) Cleaner of Vehicle/equip. (11.6) Industrial trk/tractor operator (9.9)	739	77.0%

Table 4: Employer and Occupational Mobility for Individuals working in t and t+1

(*percent*) (*row percent*) (*col. percent*)	Same Employer $t+1$	Diff. Employer $t+1$	Total
Same Occupation (10 Cat.) $t+1$	59% 84% 92%	11% 16% 31%	70%
Diff. Occupation (10 Cat.) $t+1$	5% 16% 8%	25% 84% 69%	30%

3-digit detail level is only 12% of workers. This is a reasonable figure representing mobility within firms. The fact that this number is not overstated provides reasonable assurance of the reliability of the observed occupation changes. This feature makes the NLSY97 particularly desirable for a model that looks at a finer level of occupational choice.

Table 4 provides details on the mobility of this cohort of workers. The table shows that even using the broader 10 occupation categories, workers have a 30% chance of working in a different occupation in adjacent periods. Most notable is that 17% of occupation changes at the 10 category level occur within the same firm. These occupation changes would otherwise be ignored using the previous literature's edits.

5 Results

One of the primary deliverables of the model is the estimates of the distributional parameters and correlation across occupational abilities. These estimates are found in table 5. The first row reports the standard deviation of the unconditional, marginal ability distribution for each occupation. Given that μ_{ij} enters additively in the log-wage specification, these values have a direct interpretation as the percentage difference in wages, all else equal. For example, the ability standard deviation for sales is 0.33. This means that an individual with a sales ability match at one standard deviation above zero will receive a 33% higher wage than an individual who's sales ability is at the mean. These numbers represent the potential gains from search, where an individual can potentially realize very large wage increases simply by finding an occupation they are well suited for.

The second row in table 5 contains the estimate of the covariance of schooling ability and occupational ability. Using the expression in equation (3) for the initial beliefs regarding occupational ability conditional on schooling ability, we get an interpretation for this parameter. Since $\mathbf{E}_{i1}(\mu_{ij}) = \delta_{sJ_{(j)}}\mu_{is}$, we interpret the estimates of $\delta_{sJ_{(j)}}$ as the expected match value for an individual with schooling ability at one standard deviation in the distribution. For example, an individual at one standard deviation in the schooling distribution will expect wages in professional occupations to be 16% higher due to their ability compared to an individual at the -1 standard deviation in the schooling ability (2×0.0816), holding constant the level of education.

The lower triangular matrix in table 5 is the estimated correlation matrix across occupational abilities. With the exception of one element, this matrix is uniformly positive, implying an underlying element of general ability that is persistent throughout the correlation matrix. This means that when a worker learns they are good at one thing, they are likely to be good at everything. The single exception is between schooling ability and transportation workers. The negative covariance suggests that those with high schooling ability will likely be less productive in transportation occupations, holding the level of schooling constant.

Looking across this correlation structure we detect some additional patterns that surface from the data. For example the second largest correlation, that between sales occupations and office and administrative occupations (0.7361), is between occupations often defined by the 2002 census code as a single occupation "Sales and Office Occupations". This is to say that the census has likely grouped these occupations based on the assumption that these occupations share a similar set of skill requirements. Given the estimated high correlation coefficient, the data suggests there may be some merit to this claim (although they are not perfectly correlated). The same holds true between production occupations and maintenance and repair occupations, which has the highest correlation of (0.7448).

One of the drawbacks of the raw correlation matrix is that it provides little information regarding the relative correlation between any two elements, meaning that a spurious correlation could exist between two occupations that is driven by the two occupations being correlated with a third

Table 5: Distributional Parameters and Correlation Structure [a]

	Man./ Bus.	Prof.	Service	Food	Sales	Office/ Admin.	Const- ruct.	Maint./ Repair	Prod- uction	Transp.
Ability SD $\Delta_{j,j}$	0.3280** (0.0287)	0.3385** (0.0240)	0.2293** (0.0363)	0.2705** (0.0242)	0.3329** (0.0235)	0.2873** (0.0208)	0.2777** (0.0156)	0.2846** (0.0313)	0.2659** (0.0317)	0.2352** (0.0155)
Schooling Cov. δ_{sj}	0.0780 (0.0724)	0.0816 (0.0604)	0.0193 (0.0472)	0.1143** (0.0390)	0.0506 (0.0397)	0.0920** (0.0408)	0.0186 (0.0398)	0.0703 (0.0468)	0.0340 (0.0440)	-0.0421 (0.0364)
Correlations										
Mgr/Bus.	1.0000 (0.0000)	-	-	-	-	-	-	-	-	-
Professional	0.6658** (0.1293)	1.0000 (0.0000)	-	-	-	-	-	-	-	-
Service	0.1378 (0.1866)	0.2264 (0.1974)	1.0000 (0.0000)	-	-	-	-	-	-	-
Food Service	0.3658** (0.1651)	0.2954** (0.1473)	0.6404** (0.1257)	1.0000 (0.0000)	-	-	-	-	-	-
Sales	0.5987** (0.1389)	0.5400** (0.1500)	0.4860** (0.1901)	0.4082** (0.1740)	1.0000 (0.0000)	-	-	-	-	-
Office/Admin. Support	0.7345** (0.1147)	0.4303** (0.1623)	0.3461 (0.2192)	0.4599** (0.1381)	0.7361** (0.0982)	1.0000 (0.0000)	-	-	-	-
Construction	0.5492** (0.1374)	0.1151 (0.2257)	0.3355* (0.2007)	0.3528** (0.1361)	0.4465** (0.1158)	0.6854** (0.1602)	1.0000 (0.0000)	-	-	-
Maintenance / repair	0.2911 (0.2695)	0.5335** (0.2153)	0.6573** (0.1606)	0.5213** (0.1467)	0.5834** (0.1738)	0.5137** (0.1582)	0.3354** (0.1379)	1.0000 (0.0000)	-	-
Production	0.4987** (0.2146)	0.5079** (0.2049)	0.3904** (0.1784)	0.2911* (0.1694)	0.4787** (0.1702)	0.6271** (0.1670)	0.5278** (0.1051)	0.7448** (0.1273)	1.0000 (0.0000)	-
Transportation	0.4857** (0.1922)	0.1774 (0.2057)	0.2628 (0.1984)	0.1837 (0.1841)	0.4352** (0.1773)	0.5233** (0.1489)	0.5847** (0.1023)	0.2821* (0.1476)	0.6145** (0.1348)	1.0000 (0.0000)

a Standard errors are reported in parenthesis and are constructed by 180 boot straps of the model
** Denotes significance at the 5% level.
* Denotes significance at the 10% level.

occupation. As an example, it may not be clear why sales occupations and production occupations have a correlation of (0.4787). One possibility is that the skills required are common across these two occupations, so a high ability in one implies a high ability in the other. The second possibility is that they share no skill requirements and the entire correlation is driven by the fact that sales occupations are correlated with maintenance and repair occupations, and likewise production occupations are also correlated with maintenance occupations. So any relationship we observe between sales and production is due entirely to their relationship to maintenance and repair.

To address this issue, table 6 compares the correlation of two occupations in isolation, margining out the effect of the other $J-2$ occupations. This procedure isolates the direct correlation effect and removes the secondary effects mentioned above. The most striking result from table 6 is that about one-third of the the correlation pairs are now negative. These negative numbers have the implication that it is possible for workers to learn they are good in an occupation, by actually learning they are bad at another occupation.

For example, returning to the paired correlation between sales occupations and production occupations, in table 6 it is (-0.4245). This is in contrast to the (+0.4787) found in the full correlation structure in table 5. An example of what these numbers mean is that if a worker learns they have a high ability in maintenance occupations, looking at table 5, they will infer that they have among other things, a high ability in sales occupations and production occupations. The fact that the conditional correlation between sales and production is negative means that if the worker then moves to a production occupation and discovers a poor match, then their belief about their sales match will actually increase. The idea in simple terms is that assume there are two components that make up ability for maintenance and repair occupations, one that is related to sales and one that is related to production. By only observing their maintenance and repair ability, the worker cannot disentangle which source is driving their high ability. Once they learn that it is not the production portion by gaining production experience, they then infer that it is actually the sales component, increasing their expectation of their sales ability.

The estimates of the wage parameters are presented in table 7. Given the non-linear way education and experience are modeled, the bottom half of the table contains a number of imputed

Table 6: Paired Conditional Ability Correlation Matrix

	Man./ Bus.	Prof.	Service	Food	Sales	Office/ Admin.	Const- ruct.	Maint./ Repair	Prod- uction	Transp.
Mgr/Bus.	0.0000	0.7181	-0.1754	0.2441	0.0816	0.3524	0.3457	-0.3821	0.0375	0.1810
Professional	0.7181	0.0000	-0.0130	-0.0532	0.2150	-0.2807	-0.4047	0.2839	0.2012	-0.2404
Service	-0.1754	-0.0130	0.0000	0.5105	0.2356	-0.1125	0.1527	0.2889	-0.0512	0.0725
Food Service	0.2441	-0.0532	0.5105	0.0000	-0.2166	0.0951	0.0154	0.2676	-0.3026	0.1203
Sales	0.0816	0.2150	0.2356	-0.2166	0.0000	0.5283	-0.0792	0.3390	-0.4245	0.2565
Office/Admin. Support	0.3524	-0.2807	-0.1125	0.0951	0.5283	0.0000	0.2656	0.0095	0.2641	-0.0710
Construction	0.3457	-0.4047	0.1527	0.0154	-0.0792	0.2656	0.0000	0.0419	0.1290	0.0857
Maintenance / repair	-0.3821	0.2839	0.2889	0.2676	0.3390	0.0095	0.0419	0.0000	0.6758	-0.2923
Production	0.0375	0.2012	-0.0512	-0.3026	-0.4245	0.2641	0.1290	0.6758	0.0000	0.5545
Transportation	0.1810	-0.2404	0.0725	0.1203	0.2565	-0.0710	0.0857	-0.2923	0.5545	0.0000

returns based off of the estimates. The wage estimates suggest an interesting trade-off workers may face highlighted in Miller (1984), which is between occupations with high expected wages and low variance versus low expected wages, but high variance in unknown abilities. For example, a white male high school graduate choosing their first occupation in construction can expect about 15% higher wages than if they worked in a sales occupation. However, the standard deviation in unknown occupational ability for sales occupations is about 20% higher than construction, implying their are potentially more informational benefits from sales occupations.

The structural choice estimates are in table 8. The coefficient of relative risk aversion, ρ, is estimated at 2.433. This figure represents a moderate level of risk aversion and is consistent with other findings in the literature on individual's risk preferences (e.g. Mehra and Prescott (1985)). In comparison to the entry costs, the estimated level of risk aversion implies that it is a less significant search friction. Entry costs vary across occupations. The main entry cost, α_{j2}, measures the cost to finding work in an occupation if they were not engaged in that occupation in the previous period. An additional entry cost, α_{j3}, is incurred if the worker also has no experience in that occupation. Production occupations have the highest entry costs for a new worker of all of the occupations . Given the marginal utility of income at entry level wages, the entry costs for production occupations of $(-1.952-5.193)$ is equivalent to about \$26,000, which is more than the entry level annual salary.[21] The entry costs for an individual who is returning to a production job, is much less at about \$7,000, or about 1/3 of an annual salary. This figure makes since if 4 months is the approximate time it takes to secure a job in production.

Food service on the other hand has the lowest entry costs for new workers with no food service experience, with a dollar equivalent of about \$2,000. Those returning to food service face entry costs of about \$4,400. Given that re-entry costs are double the initial entry costs implies workers are more reluctant to return to food service after leaving than they were to initially work in it. It is unclear whether these two extremes of initial entry costs (\$26,000 for production or \$2,000 for

[21] $\frac{\partial u}{\partial wage} = \alpha_w wage^{-\rho}$. With entry level wages at about \$9.20, the marginal utility of wages is 0.5425. This number is the marginal value of a \$1 increase in wages. If workers work 2,000 hours per year, then $1 util$ is equivalent to about \$3,700

Table 7: Wage Estimates and Imputed Returns to Accumulated Human Capital [a]

	Man./ Bus.	Prof.	Service	Food	Sales	Office/ Admin.	Const- ruct.	Maint./ Repair	Prod- uction	Transp.
Constant	1.4626*	0.9331	2.5066**	1.3726**	1.6366**	2.5048**	2.8881**	3.8255**	2.6864**	1.7220**
θ_{j1}	(0.8604)	(1.0727)	(0.4029)	(0.3935)	(0.5730)	(0.6075)	(0.4906)	(1.0636)	(0.4502)	(0.4348)
educ	0.0406	0.1525	-0.0927	0.0991	0.0312	-0.0884	−0.1371*	−0.2896*	-0.0944	0.0374
θ_{j2}	(0.1180)	(0.1497)	(0.0589)	(0.0626)	(0.0855)	(0.0924)	(0.0796)	(0.1617)	(0.0712)	(0.0666)
$educ^2/100$	0.1614	-0.2969	0.5349**	-0.2875	0.0964	0.4870	0.7493**	1.3209**	0.5272*	0.0093
θ_{j3}	(0.3933)	(0.5121)	(0.2241)	(0.2532)	(0.3119)	(0.3467)	(0.3317)	(0.6123)	(0.2933)	(0.2664)
$Expr_j = 0$	-0.0779	-0.0006	-0.0210	0.0501	-0.0371	-0.0287	0.0896**	-0.0032	-0.0249	-0.0087
θ_{j4}	(0.0676)	(0.0465)	(0.0533)	(0.0469)	(0.0411)	(0.0552)	(0.0353)	(0.0597)	(0.0385)	(0.0426)
$Expr_j$	0.1803**	0.1039**	0.0610	0.0330	0.0868**	0.0877*	0.0542**	0.0794	0.0658**	0.0575*
θ_{j5}	(0.0618)	(0.0476)	(0.0409)	(0.0411)	(0.0378)	(0.0514)	(0.0194)	(0.0518)	(0.0319)	(0.0342)
$Expr_j^2/100$	-1.3123	-0.5347	-0.1944	-0.1106	-0.5709	-0.2409	0.0005	-0.3041	0.0024	-0.0500
θ_{j6}	(1.0206)	(0.8465)	(0.6584)	(0.6461)	(0.6056)	(0.9198)	(0.2181)	(0.9020)	(0.5083)	(0.5150)
$Expr_{-j}$	0.0693**	0.0106	0.0401**	0.0294*	0.0698**	0.0583**	0.0175*	0.0524**	0.0193*	0.0392**
θ_{j7}	(0.0144)	(0.0152)	(0.0122)	(0.0151)	(0.0119)	(0.0093)	(0.0092)	(0.0114)	(0.0099)	(0.0088)
SD. tech. shock σ_j	0.2728**	0.2416**	0.2003**	0.2125**	0.2231**	0.2059**	0.2609**	0.2520**	0.2132**	0.2277**
	(0.0263)	(0.0162)	(0.0277)	(0.0217)	(0.0207)	(0.0181)	(0.0171)	(0.0356)	(0.0185)	(0.0157)
Imputed Returns to Accumulated Human Capital										
Base Log-Wage 12 Yrs Ed [b]	2.1818**	2.3357**	2.1646**	2.1484**	2.1500**	2.1455**	2.3212**	2.2520**	2.3131**	2.1844**
	(0.0616)	(0.0483)	(0.0350)	(0.0271)	(0.0299)	(0.0257)	(0.0305)	(0.0360)	(0.0316)	(0.0270)
Average Return per Yr. Col. [c]	0.0858**	0.0694**	0.0571**	0.0187	0.0582**	0.0480**	0.0726**	0.0802**	0.0532**	0.0400**
	(0.0174)	(0.0155)	(0.0144)	(0.0166)	(0.0112)	(0.0134)	(0.0194)	(0.0215)	(0.0195)	(0.0162)
Average Rerutn per Yr.	0.1239**	0.0962**	0.0588**	0.0469**	0.0636**	0.0874**	0.0902**	0.0795**	0.0762**	0.0665**
Exper. [d]	(0.0173)	(0.0171)	(0.0161)	(0.0147)	(0.0111)	(0.0184)	(0.0082)	(0.0230)	(0.0137)	(0.0127)

a Standard errors are reported in parenthesis and are constructed by 180 boot straps of the model
b Imputed by $\theta_{j1} + (ed = 12)\theta_{j2} + (ed = 12)^2/100\theta_{j3}$
c Imputed by $[(\theta_{j1} + (ed = 16)\theta_{j2} + (ed = 16)^2/100\theta_{j3}) - (\theta_{j1} + (ed = 12)\theta_{j2} + (ed = 12)^2/100\theta_{j3})]/4$
d Imputed by $(\theta_{j4} + (Expr_j = 4)\theta_{j5} + (Expr_j = 4)^2/100\theta_{j6})/4$
** Denotes significance at the 5% level.
* Denotes significance at the 10% level.

Table 8: Structural Choice Estimates[a]

Employment

	Manager/ Business	Profes- sional	Other Service	Food Service	Sales
Non-Pecuniary Constant, α_{j1}	5.930 (4.545)	6.462^{*} (3.845)	8.441^{**} (4.263)	4.012 (4.247)	9.590^{**} (4.262)
Entry Cost ($d_{t-1} \neq j$), α_{j2}	-2.922^{**} (0.382)	-2.059^{**} (0.286)	-2.329^{**} (0.231)	-1.199^{**} (0.187)	-2.097^{**} (0.205)
No Experience ($Expr_j = 0$), α_{j3}	-0.119 (1.543)	-3.025^{**} (1.509)	-1.287 (0.917)	0.635 (1.025)	-1.708^{*} (0.965)

	Office/ Admin.	Cosntruc- tion	Mainten- ance	Produc- tion	Transpor- tation
Non-Pecuniary Constant, α_{j1}	9.118^{**} (4.281)	8.739^{**} (4.068)	13.636^{**} (4.271)	13.878^{**} (4.604)	10.488^{**} (4.177)
Entry Cost ($d_{t-1} \neq j$), α_{j2}	-1.635^{**} (0.180)	-1.423^{**} (0.163)	-2.441^{**} (0.262)	-1.952^{**} (0.208)	-1.766^{**} (0.177)
No Experience ($Expr_j = 0$), α_{j3}	-2.552^{**} (0.943)	-3.179^{**} (0.962)	-3.997^{**} (1.258)	-5.193^{**} (1.123)	-1.189 (0.992)

	Utility Over Income and Risk[b]
Coef. on Wages, α_w	120.031 [7.350 , 1797.669]
Rel. Risk Aversion, ρ	2.433 [1.285 , 3.806]

Schooling

Constant, α_{s1}	4.453^{**} (0.924)
College Attendance, α_{s2}	-3.828^{**} (0.754)
Re-entry HS, α_{s3}	-3.519^{**} (0.200)
Re-entry Col., α_{s4}	-1.289^{**} (0.106)
Schooling Ability, α_{s5}	0.751^{**} (0.243)

[a] β is fixed to 0.95. Standard errors are reported in parenthesis and are constructed by 180 bootstraps of the model.

[b] Bracketed numbers represent the 95% confidence interval of the parameter estimate from the 180 bootstraps.

** Denotes significance at the 5% level.

* Denotes significance at the 10% level.

food service) are reasonable in comparison to the previous literature's estimates given that these occupations are typically aggregated together into blue collar occupations.

Entry costs may be more or less relevant in the context of the occupation specific non-pecuniary constant, α_{j1}. Even though production has the largest entry costs, if also has the largest non-pecuniary benefit, leaving it on net the occupation with the third highest (behind transportation and maintenance) non-pecuniary benefit for new entrants (13.878-1.952-5.193). This very large constant means that once workers have experience in production, and in particular worked in production in the previous period, they are very likely to work in production in the next period. This pressure to continue in the occupation from the previous period functions as a major friction to sorting on ability, where even workers who receive poor information on their ability will be reluctant to seek out a better match.

Turing to the results on schooling preferences, on average individuals receive a positive utility for attending school, with a substantial reduction in utility for attending college. At median, entry level wages, with marginal utility of wages equal to 0.5425, the disutility of college of -3.828 implies a cost of about $14,000 per year of college. Schooling ability significantly effects the decision to attend school. The coefficient estimate of 0.751 implies that an individual at plus one standard deviation in the schooling ability distribution receives the equivalent of a $2,800 tuition subsidy compared to an individual at the mean.

6 Analysis: Occupational Choice, Sorting on Ability, and Wages

This section explores in greater detail the importance of sorting on ability in observed occupational choices. The large variances of the occupational ability distribution indicate that workers have much to gain about learning where they fall in each distribution. However, in reality, there are a myriad of factors in addition to sorting on ability that drive occupational choice. Using the parameter estimates, we can simulate career paths to gain much needed insight into the aptitude of workers to sort on ability and perhaps more importantly the effect that sorting on ability has on wage growth.

We can define the individuals expected *ex-ante* occupational match as,

$$\bar{\mu}_{it} = \frac{\sum_{j=1}^{J} \Omega_j(i,t)\mu_{ij}}{\sum_{j=1}^{J} \Omega_j(i,t)} \tag{44}$$

where $\Omega_j(i,t)$ represents the individual's optimal choice strategy given the state at time t. This measure has a nice interpretation in that as it moves away from their average match, it implies that the worker is choosing a match with certainty. For example if the *ex-ante* expected ability equals their maximum (or minimum) ability, then we can say that they are choosing their maximum (or minimum) ability with certainty.

The empirical model is robust in the sense that it allows many other factors not relating to ability to influence occupational choice (e.g. entry costs and risk aversion), meaning that the model does not require people to sort on ability. Therefore, on an individual level, if we see a worker choose their maximum occupation with certainty we cannot say for sure that this is driven by sorting on ability because this outcome could be driven by outside factors. [22] However, the theoretical ability matching model, in the absence of search frictions, says that the fraction of workers whose *ex-ante* expected ability equals their maximum ability should continually improve over time. If we observe this trend in the data on an aggregate level, then this suggests that sorting on ability is relevant to occupational choice. However, if other factors dominate the workers aptitude to sort on ability, then there will likely to no discernible patterns in the *ex-ante* expected ability.

To see what trends (if any) exist in the data I simulate 1 million careers using the parameter estimates. Rather than looking at the full distribution of abilities, for each worker their set of ten ability matches are grouped into three categories: O_{1-3} corresponds to the range of the individual's top three occupational abilities, O_{7-10} corresponds to the range of the individual's worst three occupational abilities, and the remaining category, O_{4-6}, includes the ability levels that are not in the best three or worst three. Table 9 tracks the fraction of workers who's *ex-ante* expected

[22]For example, they could have learned that this is their highest ability and are thus sorting into that occupation. Alternatively, they could be a high school drop out, and perhaps it is a common rule for all high school drop outs to go into transportation occupations, of which this individual happens to have their highest match in transportation. Another possibility is that workers always choose occupations at random and then stay in that job forever. In this workers case that just happened to be their high ability occupation. The first explanation entails the worker sorting on ability, while the later two explanations are driven by strictly random search processes.

ability falls in each segment of the distribution of their abilities. The analysis is broken down by educational attainment because these workers have very different priors, occupational search patterns, and labor market attachment.

It is important to understand what table 9 is measuring. It is the *ex-ante* expected ability, not the *ex-post*. Therefore, the column $\Pr(\bar{\mu} \in O_{7-10})$ does not measure the fraction of workers who choose an occupation in their bottom three ability *ex-post*, it is the fraction who are likely choose one of their worst occupations *ex-ante*. To draw the distinction, assume workers choose occupations completely randomly (i.e. with probability of choosing each occupation, $1/J$). Then *ex-post*, 30% of workers will end up in one of their three worst occupations. This fraction is not very informative because these workers ended up in these occupations completely unintentionally. The measure we are interested in, $\Pr(\bar{\mu} \in O_{7-10})$, is the fraction of workers who intentionally, or in some sense, are likely to choose one of their worst three occupations. In the context of the random decision model, this value is zero because the randomness implies that *ex-ante* their match is likely in the middle of the distribution. Looking at *ex-ante* expected ability is a way to separate intentional and unintentional occupational sorting on ability.

The *ex-ante* expected match at zeros years of labor market experience is quite different across educational categories. The table shows that nearly all (96.6%) of high school gradates with zero years of labor market experience have an expected match in the middle range of their match values. It is not surprising that the overwhelming fraction of high school entrants have an expected ability in the middle of their ability distribution since they are new entrants. However, given initial choices, 2.4% will end up in one of their top three occupations, and 1.0% end up in one of their bottom occupations. It is also not surprising that these fractions are not equal and biased toward the upper end of the distribution. What this means is that high school graduates have priors that will on average put them in a higher ability occupation. Therefore, given however high school graduates choose occupations, there are 2.4% that are likely ending up in one of their best matches and only 1.0% ending up in a poor match.

However, given that these are new entrants into the labor force, it is likely that the large pooling in the middle of the distribution is due to the fact that sorting on ability at this stage of the career

Table 9: Occupational Sorting by Labor Market Experience and Education

Education Group	Labor Market Experience (Yrs.)	$\Pr(\bar{\mu} \in O_{7-10})^a$	$\Pr(\bar{\mu} \in O_{4-6})$	$\Pr(\bar{\mu} \in O_{1-3})^b$
Less Than HS	0	1.9%	86.9%	11.2%
	1	7.4%	72.9%	19.7%
	2	7.6%	71.4%	21.0%
	3	7.6%	70.6%	21.8%
	4	7.6%	69.6%	22.7%
	5	7.8%	68.9%	23.3%
	6	7.8%	68.0%	24.2%
	7	8.0%	67.1%	24.9%
	8	8.4%	65.9%	25.7%
HS	0	1.0%	96.6%	2.4%
	1	8.8%	77.3%	13.9%
	2	10.0%	73.2%	16.8%
	3	10.8%	70.3%	18.9%
	4	11.3%	68.1%	20.7%
	5	11.8%	66.1%	22.2%
	6	12.2%	64.1%	23.7%
	7	12.5%	62.4%	25.1%
	8	12.6%	61.4%	26.0%
	9	13.2%	59.5%	27.4%
College	0	8.9%	74.8%	16.3%
	1	14.8%	61.4%	23.8%
	2	16.2%	58.0%	25.8%
	3	16.7%	56.3%	27.0%
	4	16.6%	55.6%	27.8%
	5	16.1%	55.3%	28.6%

[a] The fraction of individuals likely to choose one of their worst three occupational abilities.
[b] The fraction of individuals likely to choose one of their top three occupational abilities.

is more or less unintentional (random).

The difference between 2.4% and 1.0% is likely inconsequential. However, for high school drop outs we see the same pattern, but on a much larger and more convincing scale. Given their optimal choices, 11.2% have an expected match that falls within the range of their top three occupational matches prior to any labor market experience. This is evidence that for these workers, their priors really matter. Given their knowledge of their schooling ability and correlation structure, they are able to narrow down which occupations they are likely to have high (or low) ability and most importantly showsstrong evidence that they are sorting (intentionally) on this information. Conversely, only 1.9% have an expected match in the range of their bottom three abilities.

College graduates show a similarly strong pattern in their initial priors, with 16.3% likely to choose one of their top three occupational abilities, even having no labor market experience. What is interesting is that this number is only twice as large as the fraction of college graduates who are likely to choose one of their worst three occupations (compared to 6 times as large for high school drop-outs). What's driving these 8.9% of college graduates to choose one of their three worst occupations is unclear. Two possible explanations are that college graduates tend to choose occupations with higher variances in abilities (managers/ business operations and professionals), or that they knowingly have a poor match but choose the occupation for other reasons (e.g. they provide high returns to education).

The table tracks the changes in this distribution across labor market experience. The largest change in these percentages occurs between zero years and one year of labor market experience. With one year experience, individual choices become much more heterogeneous. For high school graduates, 8.8% will likely choose one of their three worst occupational matches, up from 1.0% when they have zero years labor market experience. What drives this large increase is the switching costs in the model. For workers with no labor market experience that randomly initially choose one of their worst occupations, the switching costs in the model suggest that these workers are likely to stay in this occupation despite the low match. Given that these workers are likely to choose this poorly matched occupation again, their *ex-ante* expected ability will fall accordingly and the fraction of workers who are likely to work in one of their three worst occupations will increase.

For high school graduates we observed a 7.8 point increase in the fraction of workers in one of their worst matches between zero and one year experience. For these same workers we observe a 11.5 point increase in the fraction who are likely to work in one of their best occupations. This asymmetric increase gives positive evidence that, although workers face switching costs, sorting on ability has a dominant role in occupational choice. Looking further down the career with 9 years labor market experience, 27.4% of high school graduates are likely to work in one of their top three occupations. Comparatively only 13.2% are likely to work in one of their worst three. This strongly indicates that sorting on ability plays an important role in occupational decisions.

High school graduates show the largest gain in the fraction expecting to work in one of their top three occupations over the initial stages of their career. A similar pattern exists for high school drop outs and college graduates, however, since these workers relied more heavily on their priors, the gains from learning in the labor market where not as large. The more asymmetric these distributions become over time, implies a greater importance of intentional sorting on ability in making career decisions. The asymmetry can be measured as the fraction of workers likely to work in one of their top three occupations versus the fraction likely to work in one of their bottom three. By this measure, sorting on ability is most important for high school drop outs. At five years experience these workers are nearly three times more likely to be working in one of their top three occupations compared to one of their bottom three occupations. Comparatively, high school gradates are twice as likely and college gradates are 1.75 times as likely. This high fraction of college graduates in low matches (16.1%) suggests that additional factors play an important role for occupational decisions of college graduates other than sorting on ability. One attributing factor may be that college graduates may choose a particular occupation not because they have a high ability in that occupation, but because the returns to college are high in that occupation. If they exercise flexibility in occupational choice, they may forgo some of the returns to education.

To assess the wage gains associated with ability sorting, I will only focus on the match component of wages, μ_{ij}, which measures the percentage contribution of ability to wages. We are interested in looking at how the average match value evolves over the career. Table 10 shows the expected value of μ_{ij} at different points in the career for high school graduates. Given the optimal

Table 10: Wage Gains (in percent) Due to Sorting on Ability For High School Graduates

Labor Market Experience (Yrs.)	$\mathbb{E}(\mu_{ij})$ (a)	$\mathbb{E}(\mu_{ij})$ (working) (b)	$\mathbb{E}(\mu_{ij})$ (2^{nd} best, can exit labor force) (c)	$\mathbb{E}(\mu_{ij})$ (2^{nd} best, can't exit labor force) (d)
0	-0.1%	0.5%	0.3%	1.1%
1	2.4%	3.5%	1.8%	1.1%
2	2.9%	4.0%	2.2%	1.2%
3	3.3%	4.4%	2.5%	1.3%
4	3.9%	4.9%	2.8%	1.6%
5	4.4%	5.4%	3.2%	2.0%
6	5.2%	6.1%	3.9%	2.7%
7	6.1%	6.8%	4.5%	3.4%
8	6.5%	7.1%	4.8%	3.8%
9	7.5%	8.0%	5.6%	4.7%

decisions rules $\Omega_j(i,t)$, which is the probability that individual individual i chooses occupation j, given their information at t, the expected ability in the distribution is calculated as,

$$\mathbb{E}(\mu_{ij}) = \frac{\sum_i \sum_j \Omega_j(i,t)\mu_{ij}}{\sum_i \sum_j \Omega_j(i,t)} \tag{45}$$

Table 9 indicated that priors beliefs did not likely play a role in initial occupational decisions for high school graduates. As we can see in table 10 the expected match value for high school graduates with zero years labor market experience is essentially zero, providing further evidence that these high school graduates possess little information to sort on ability. However, as their careers' progress, the average ability increases steadily by 0.5% to 1.0% per year such that with nine years labor market experience, on average, 7.5% of wages is due to sorting on ability for high school graduates. Given that table 9 shows that 13.2% of high school graduates are in one of their worst three matches, there is likely a lot of heterogeneity underlying this average ability. However, an average increase in wages of 7.5% is a sizable contribution to wage growth and for most occupations greater than the return to one year of college.

A second important aspect of the model is to look at how individuals second best options are evolving over time. The correlated learning structure suggests as workers become more informed about their ability in their current occupation, they will likely have more information about their abilities overall. Specifically, if workers are forced out of their number one occupational choice, they will be able to compensate for this loss using the information they have acquired. This is in contrast to the results generated from the independence assumption, where a job-loss results in the destruction of the entire 7.5% wage gains.

To analyze if individuals' second best choices are improving with labor market experience, we will look at the difference in the expected match for those who are employed, who are forced out of their preferred occupation and forced into an alternative occupation. These values are reported for high school graduates in table 10. The expected match for high school graduates conditional on working reported in column (b) is slightly higher at all levels of labor market experience compared to the population average because those who actually choose employment are likely to do so because they have on average a higher ability. The column (c) shows the expected match if these same workers were forced out of their first choice occupation into a second choice. One option for these workers is to exit the labor force. Column (c) reports the expect ability allowing some workers to exit the labor force. While the second best choice for these workers is almost always less than the expected match of their primary occupation, the expected ability in the second best occupations is able to keep pace with the growth in the expected ability for the primary occupation. Rather than losing the entire average 8.0% for employed high school graduates with 9 years labor market experience, if displaced, these workers will move into occupations with an expected ability of 5.6%, implying only a moderate loss of wage due to lower matched occupation of -2.4%.

Column (d) shows the expected ability if all of these workers are forced to find new occupations (i.e. workers are not allowed to exit the workforce). The expected ability of this population is lower than the case when we allow for exits, implying that the individuals that are most likely to exit the labor force do so because they do not have a suitable second best option. While this number is lower, it is still quite large, where workers are able to recover a sizable amount of their loss in wages due to matching through correlated learning.

Table 11 shows the wage gains for the other education groups. College graduates appear to have the highest expected ability match, contributing to 9.1% of wages at five years of labor market

Table 11: Wage Gains (in percent) Due to Sorting on Ability by Education

	$\mathbb{E}(\mu_{ij})$		
Labor Market Experience (Yrs.)	Less than HS	High School	College Grad.
0	-3.7%	-0.1%	7.1%
1	-1.6%	2.4%	8.8%
2	-0.9%	2.9%	8.4%
3	-0.3%	3.3%	8.3%
4	0.4%	3.9%	8.5%
5	1.1%	4.4%	9.1%
6	2.0%	5.2%	–
7	3.0%	6.1%	–
8	4.0%	6.5%	–
9	–	7.5%	–
Random Match Values	-5.6%	0%	3.9%

[a] Labor market experience for high school drop outs begins at age 16, age 18 for high school graduates, age 19 for some college, and age 22 for college graduates.

[b] The probability that an individual with a given level of education and labor market experience chooses an occupation possessing one of their worst three abilities.

[c] The probability that an individual with a given level of education and labor market experience chooses an occupation possessing one of their top three abilities.

experience. However this number does not truly reflect the gains of sorting. Given the positive correlation structure, college graduates are likely to be much better in any of the occupations. We would expect their average match value to be positive regardless of the decision rule they used. The bottom row of table 11 shows the expected match value if individuals chose occupations with equal probability. For high school graduates the expected ability at zero years experience is identical to the expected match if individuals randomly choose occupations, suggesting in some sense again that these workers have little information to base their initial choices on.

College graduates on the other hand have an expected match of of 3.9% if they choose occupations at random, implying an average wage gain of 5.2% (9.1-3.9) due to occupational sorting. What is perhaps most interesting is that the majority of these gains, 3.2%, are realized before these workers have any labor market experience and only achieve an additional 2% on average once in the labor market. This means that for college graduates, priors play a very important role in sorting on ability.

High school drop-outs show a similar story. If this group where to choose occupations randomly, this would imply an expected occupational match of -5.6%, suggesting that these workers are in general, lower ability. However, given their prior information, they are able to improve their position in the labor market by an average of about 2.0% percentage points. Although not as large as college graduates, initial sorting is relevant for high school drop outs. After nine years labor market experience high school dropouts have an expected ability value of 4.0% implying a gain on average of 9.6% $(4.0 + 5.6)$.

The collection of results from the simulations begin to describe in detail how sorting on ability effects occupational choices and wage growth. In general the results show strong evidence that workers sort into occupations where they have higher ability as they learn through labor market experience. These results vary widely across educational groups, where high school drops out appear to the be most likely to move out of poorly match occupations. College graduates on the other hand are much more likely to stay in occupations where they have very low ability. The fact that either group would choose to stay in a poorly matched occupation highlights the role of other non-pecuniary factors which drive occupational decisions in addition to sorting on ability. The exact factors leading to the asymmetric sorting on ability by education level remains for future study. As high school drop outs are the most likely to sort on ability they are able to increase wages by 9.5% with 9 years labor market experience. High school graduates and college graduates increase wages 7.5% and 5.2% respectively.

7 Conclusion

This paper develops and estimates an individual model of occupational choice and learning that allows for correlated learning across occupation specific abilities. Flexibly allowing information to enter the worker's mobility decision in this manner broadens the type of career moves that occupational matching models can address. Specifically, the model is able to capture both lateral occupational moves and vertical occupation moves, where high ability individuals change careers due to promotion. More generally, the model not only addresses what drives individuals to change careers, but also how information effects their decision of what new occupation to go to.

Endogenizing information in this way significantly increases the computational burden. I address these empirical challenges by utilizing the Expectation and Maximization (EM) algorithm to uncover the persist, high dimensional unobservables in the data. This approach is particularly well suited for the learning framework as it breaks the curse of dimensionality so the computational complexity only grows linearly in the number of occupations. The empirical strategy likely has broad applicability to other learning models.

The model is estimated on the National Longitudinal Survey of Youth 1997. The parameter estimates suggest that workers can potentially realize large wage gains by finding occupations they are well suited for. The model allows for heterogenous priors over the distribution of ability. These priors are strongly related to educational attainment, which plays an important role in driving initial occupational choices. Using the parameter estimates I simulate careers to gain insight into the importance of sorting on ability in occupational decisions and it's effect on wage growth. The simulations suggest that workers willingness to sort on ability is highly related to educational attainment, where high school drop-outs are most likely to choose occupations based on ability, and college graduates are the least likely to sort on ability. This finding provides some understanding as to why college graduates are observed in the data to be less likely to change occupations compared to other cohorts.

References

D. Ackerberg. Advertising, learning, and consumer choice in experience good markets: an empirical examination. *International Economic Review*, 44(3):1007–1040, August 2003.

K. Antonovics and L. Golan. Experimentation and job choice. Working Paper, March 2011.

P. Arcidiacono and R. A. Miller. Ccp estimation of dynamic discrete choice models. Duke University, July 2010.

C. Belzil and J. Hansen. Unobserved ability and the return to schooling. *Econometrica*, 70(5): 2075–2091, Sept. 2002.

G. Crawford and M. Shum. Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173, July 2005.

M. H. Degroot. *Optimal Statistical Decisions*. McGraw Hill, 1970.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

M. J. Dickstein. Efficient provision of experience goods: Evidence from antidepressant choice. Stanford University, August 2011.

V. J. Hotz and R. A. Miller. Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, 60(3):497–529, Jul. 1993.

W. R. Johnson. A theory of job shopping. *The Quarterly Journal of Economics*, 92(2):261–278, May 1978.

B. Jovanovic. Job matching and the theory of turnover. *The Journal of Political Economy*, 87(5): 972–990, Oct. 1979.

G. Kambourov and I. Manovskii. Rising occupational and industry mobility in the united states: 1968-97. *International Economic Review*, 49(1):41–79, February 2008.

G. Kambourov and I. Manovskii. Occupational specificity of human capital. *International Economic Review*, 50(1):63–115, February 2009.

M. Keane and K. Wolpin. The solution and estiamtion of discrete choice dynamic programming models by simulation and interpolation: Monte carlo evidence. *The Review of Economics and Statistics*, 76(4):648–672, Nov. 1994.

M. Keane and K. Wolpin. The career decisions of young men. *Journal of Political Economy*, 105 (3):473–522, June 1997.

R. Mehra and E. C. Prescott. The equity premium puzzle. *Journal of Monetary Economics*, 15(2): 145–161, Mar. 1985.

R. Miller. Job matching and occupational choice. *The Journal of Political Economy*, 92(6):1086–1120, Dec. 1984.

D. Neal. The complexity of job mobility among young men. *Journal of Labor Economics*, 17(2): 237–261, Apr. 1999.

R. Pavan. Career choice and wage growth. University of Rochester, July 2009.

J. Rust. Optimal replacement of gmc bus engines: An empirical model of harold zurcher. *Econometrica*, 55(5):999–1033, Sep. 1987.

C. Sanders. Skill uncertainty, skill accumulation, and occupational choice. Washington University, St. Louis, December 2010.

K. L. Shaw. Occupational change, employer change, and the transferability of skills. *Southern Economic Journal*, 53(3):702–719, Jan. 1987.

P. Sullivan. A dynamic analysis of educational attainment, occupational choice, and job search. *International Economic Review*, 51(1):289–317, February 2010.

R. H. Topel and M. P. Ward. Job mobility and the careers of young men. *The Quarterly Journal of Economics*, 107(2):439–479, May 1992.

S. Yamaguchi. Tasks and heterogeneous human capital. McMaster University, June 2010.

A Mispecification Bias: Aggregating Occupations

This section discusses the mispecification bias in the ability matching model if occupations are coarsely aggregated into blue collar and white collar as is typically done in the occupational choice literature. When occupations are aggregated, this assumes that the ability vectors are perfectly correlated. To the degree that this is not true, many of the structural parameters will be biased. For example, the returns to tenure will be overstated. If workers are actually searching among many occupations, which the econometrician assumes is only one occupation, then the returns to search will be perceived as returns to work experience. In a similar way, the average match quality will be upward bias. Since people are selecting their highest match among multiple occupations, which the econometrician assumes are one.

Less obvious will be it's effect on the variance parameters of the model. Inappropriately aggregating occupations will downward bias the variance on occupational match quality. The estimator will compensate for the downward bias in the ability variance by upwardly biasing the variance of the technology shock. This bias can be demonstrated with a simple stylized model.

Assume their are two occupations (1,2) that the econometrician aggregates into a single occupation. Each worker observes one ability signal from each occupation parameterized as follows for algebraic ease.

$$\begin{aligned} w_{i1} &= \mu_{i1} + \varepsilon_{i1}, \quad \mu_{i1} \sim N(0, \sigma^2), \varepsilon_{i1} \sim N(0, \sigma^2) \\ w_{i2} &= \mu_{i2} + \varepsilon_{i2}, \quad \mu_{i2} \sim N(0, \sigma^2), \varepsilon_{i2} \sim N(0, \sigma^2) \\ &\qquad\qquad\quad \text{and } E(\mu_{i1}\mu_{i2}) = \rho \end{aligned}$$

If the econometrician assumes both wage signals are from a single match $\tilde{\mu}_i$, the posterior beliefs will be.

$$\begin{aligned} E_i(\tilde{\mu}_i) &= \left((\sigma^2)^{-1} + (\sigma^2)^{-1}\right)^{-1} \left((\sigma^2)^{-1}0 + (\sigma^2)^{-1}(w_{i1} + w_{i2})\right) \\ &= \frac{w_{i1} + w_{i2}}{3} \\ V_i(\tilde{\mu}_i) &= \left((\sigma^2)^{-1} + (\sigma^2)^{-1}\right)^{-1} \\ &= \frac{\sigma^2}{3} \end{aligned}$$

Given these values for the population, we can derive the population variance of match quality

$$
\begin{aligned}
V(\tilde{\mu}) =& E(\tilde{\mu}^2) \\
=& E\left[V_i(\tilde{\mu}_i) + E_i(\tilde{\mu}_i)^2\right] \\
=& E\left[\frac{\sigma^2}{3} + \left(\frac{w_{i1} + w_{i2}}{3}\right)^2\right] \\
=& E\left[\frac{\sigma^2}{3} + \frac{w_{i1}^2}{9} + \frac{2w_{i1}w_{i2}}{9} + \frac{w_{i2}^2}{9}\right] \\
=& \frac{\sigma^2}{3} + \frac{2\sigma^2}{9} + \frac{2\rho}{9} + \frac{2\sigma^2}{9} \\
=& \frac{7\sigma^2}{9} + \frac{2\rho}{9} < \sigma^2 \quad \text{if } \rho < \sigma^2 \text{ or } CORR(\mu_{i1}, \mu_{i2}) \neq 1
\end{aligned}
$$

Turning to the bias on the technology shock, since the total variance of wages is $V(w) = 2\sigma$, subtracting the estimate for $V(\tilde{\mu})$, $V(\varepsilon)$ is,

$$V(\varepsilon) = 2\sigma^2 - \frac{7\sigma^2}{9} - \frac{2\rho}{9} > \sigma^2 \quad \text{if } CORR(\mu_{i1}, \mu_{i2}) \neq 1$$

Given the high value of understanding sources of wage variation, the distribution of abilities and the selection problems stated earlier, these sorts of biases are extremely undesirable. Therefore, it is important that we push the empirical model to accommodate as many occupations as possible.