Selective Migration, Occupational Choice, and the Wage Returns to College Majors*

Preliminary and Incomplete

Tyler Ransom†

Duke University

September 14, 2016

Abstract

I examine the extent to which the returns to college majors are impacted by selective occupational choice and migration across locations in the U.S. There are large differences across locations in major-specific earnings premiums, migration rates, and propensity to work in related occupations. These spatial differences indicate that selective migration and occupational choice might be an important reason for cross-major earnings variation. To quantify the role of selection, I develop and estimate an extended Roy model of migration, occupational choice, and earnings where individuals choose in which U.S. state to live and in which occupation to work upon completing their education. In order to estimate this high-dimensional choice model, I make use of machine learning methods that allow for model selection and estimation simultaneously in a non-parametric setting. I find that selection in location and occupational choice is an important determinant of earnings differences across majors, but that the magnitude is small. This finding suggests that nonpecuniary preferences associated with post-graduation decisions are important determinants of labor market outcomes.

JEL Classification: I2, J3, R1

Keywords: College major, migration, occupation, Roy model

*I would like to thank Peter Arcidiacono, Esteban Aucejo, Arnaud Maurel, and seminar participants at Cal State Fullerton, Duke Economics, Duke SSRI, and Oklahoma State for their helpful discussions and comments. Special thanks to Jamin Speer for generously providing helpful code for classifying college majors in the ACS. All errors are my own.

†Contact: Social Science Research Institute, Duke University, Box 90989, Durham, NC 27708-0989. E-mail: *tyler.ransom@duke.edu*

1 Introduction

There are large differences in earnings among individuals possessing college degrees in different majors. For example in the American Community Survey, STEM and business majors each earn approximately 30% more than education majors, even after controlling for a rich set of demographic characteristics. However, there are two potential determinants of earnings differences across majors that have not been fully exmined by the existing literature: post-college migration, and post-college occupational choice.[1]

Indeed, there is an important interaction among earnings, locational choice, and occupational choice of college graduates that has yet to be fully understood. For example, education majors have the lowest earnings on average, but also have the lowest propensity to live outside their state of birth and the highest propensity to work in an occupation related to their major. On the other hand, STEM majors have the highest earnings, the highest propensity to migrate, and work in related occupations at a rate close to education majors. As a third example, there also appears to be a local labor market component to college major earnings. Business and economics majors have earnings and related occupation employment propensities that are close to STEM majors, but have migration behavior that more closely resembles education majors. Each of these three cases motivates a deeper understanding of the role of post-college decisions on the earnings of college majors.

The object of this paper is to uncover the extent to which selection into residence location and occupation biases the observed earnings differences across college majors. The Roy (1951) model is the canonical lens through which to understand selectivity bias in observed earnings, and this paper adds to the vast literature that has used that model to empirically disentangle the returns to education from selection.

As it relates to the college major question, selective migration and occupational choice could play an important role in explaining wage differences because individuals who have moved locations or who are working in a particular occupation may have made the decision in response to a more favorable wage draw. If this is the case on average, then one would expect the returns to major to be upward biased. At the same time, individuals may have non-wage preferences associated with a particular location or occupation. If these preferences outweigh the responsiveness to wage draws, then the returns to major might be downward biased.

The exact magnitude and direction of selection bias is an empirical question. To account for all possibilities, I estimate an extended Roy model that allows for non-pecuniary tastes in both the location and occupation dimensions. This paper bridges together previous work that has examined the role of selective migration on the college wage premium (Dahl, 2002; Bayer, Khan, and Timmins, 2011) and the role of selective occupational choice on the returns to college major (Kinsler and Pavan, 2015).

[1] Important exceptions to this include Altonji, Arcidiacono, and Maurel (2016a), Altonji, Kahn, and Speer (2016b), Kinsler and Pavan (2015), and Winters (Forthcoming).

Estimation of an extended Roy model is difficult when one allows for many locations and non-pecuniary preferences. To estimate the model, I supplement methods pioneered by Lee (1983) and Dahl (2002) with machine learning methods that are becoming more popular in economics (Varian, 2014; Bajari et al., 2015). The Dahl approach shows that selection can be corrected for by including a polynomial function of a small number of observed choice probabilities.[2] This polynomial serves as a multidimensional analog of the inverse Mill's ratio in the classic Heckman (1979) correction model. As a result, the researcher can obtain unbiased and consistent estimates of the selection-corrected returns using OLS.

I estimate the probabilities that enter the selection correction term using nonparametric machine learning methods, which focus on balancing in- and out-of-sample fit and allow for model selection by cross-validation. In general, machine learning algorithms are tailored towards predictive accuracy instead of causal inference and hence can predict more accurately than estimators traditionally found in the economics literature. The specific prediction setting in this paper is multiclass classification using decision trees. I classify individuals into occupations and destination locations based on their observable characteristics, which include birth location, completed college major, and demographic characteristics. The probability of belonging to a particular class (i.e. the probability of choosing a particular location-occupation combination) serves as the basis of the selection correction estimator.

I find that the returns to college major are *downward* biased by selective migration and occupational choice, but that the magnitude of the bias is small—on the order of 15%. This suggests that nonpecuniary preferences associated with location and occupation are important determinants of earnings differences across majors. This finding is consistent with Kinsler and Pavan (2015), who show that post-college occupational choice does not narrow the STEM earnings gap.

The remainder of the paper is organized as follows: Section 2 details the Roy model which serves as the empirical basis of understanding selection. Section3 outlines the statistical framework that allows me to reduce the dimensionality of the choice set. Section 4 describes the data construction and key variables used in the estimation, and Section 5 discusses the estimation of the model, including the non-parametric machine learning decision tree algorithm. Section 6 discusses the main empirical findings, and Section 7 concludes.

2 A Roy Model of Migration, Occupation, and Earnings

In this section, I introduce an extended Roy (1951) model of college major, occupation, and location choice, using the framework developed in Dahl (2002). The model is an extended Roy model because it extends Roy's original model in two ways: (i) both pecuniary and non-pecuniary factors influence an individual's decision; and (ii) there are more than two alternatives

[2] This is referred to as the index sufficiency assumption.

in the choice set.[3]

An extended Roy model serves as an appropriate lens through which to view the location decisions of college graduates because location has been shown to be an increasingly important determinant of labor market outcomes (Moretti, 2012; Diamond, 2016). Furthermore, any study of migration needs to incorporate utility maximization rather than income maximization because non-pecuniary factors such as amenities and distance are important determinants of migration decisions (Kennan and Walker, 2011; Ransom, 2016; Zabek, 2016).

2.1 Model

This section formalizes each component of the Roy model and how each component interacts with each other. The primary components of the model are earnings (the outcome equation) and preferences (the selection equation).

The framework of the model is as follows. A geographical area (e.g. the United States) is divided into L mutually exclusive locations (e.g. states). The model has two periods. In the first period, individuals are born and make human capital investment decisions. In the second period, individuals choose where to live and in which occupation to work, and receive utility from both earnings and non-pecuniary aspects of the chosen location and occupation.

The focus of this paper is on how selective migration and occupational choice in the United States affects the measured returns to the specific human capital investment of college major. A number of studies have established that earnings differentials exist across majors (e.g. Arcidiacono, 2004; Altonji et al., 2016b,a), and that certain majors have more distinct occupational distributions than others (e.g. Ransom and Phipps, 2016). This paper serves to examine how much of the cross-location earnings differentials reflect selection on location and occupation, emphasizing the fact that some locations are more conducive to certain occupations.

2.1.1 Earnings

Log annual earnings for individual i residing in location ℓ and working in occupation k are given by the following equation:

$$w_{i\ell k} = \gamma_{0\ell} + x_i\gamma_{1\ell} + s_i\gamma_{2\ell k} + \eta_{i\ell k}, \quad \ell = 1, \ldots, L \ , \ k = 1, \ldots, K \tag{2.1}$$

where x_i is a vector of individual characteristics and s_i is a $(S+4)$-dimensional vector of dummy variables indicating i's educational attainment. Specifically, the vector distinguishes among S separate college majors if i holds at least a bachelor's degree.[4] Importantly, $\eta_{i\ell k}$ will not gener-

[3] See Heckman and Taber (2008) for an overview of the original Roy (1951) model and its various extensions. Heckman and Honoré (1990) discusses identification of the Roy model, including the assumptions on the distribution of earnings that are required to generate empirical content of the Roy model.

[4] The complete set of categories is: high school dropout, high school graduate (or GED recipient), some college, college graduate (each of K major categories), and advanced degree.

ally be mean-zero because of selection, which would bias OLS estimates of $\gamma_{1\ell}$ and $\gamma_{2\ell k}$. The parameter of interest in (2.1) is $\gamma_{2\ell k}$, which measures the link between earnings, schooling investments, and location and occupational choice.

It is important to note that i's location of birth does not enter (2.1). This is because I utilize birth location as an exclusion restriction in order to separately identify non-pecuniary preferences from earnings.[5] This assumption is discussed in more detail in the following footnote.[6]

2.1.2 Preferences

Individuals have preferences for both earnings and non-pecuniary utility:

$$V_{ij\ell k} = w_{i\ell k} + u_{ij\ell k}, \quad \ell = 1, \ldots, L \ , \ k = 1, \ldots, K \tag{2.2}$$

where j indexes birth location, ℓ indexes current location, and k indexes occupation. $u_{ij\ell k}$ encompasses all non-pecuniary utility components that could determine the utility of residing in location ℓ and working in occupation k given origin j. These include location characteristics such as climate, crime, commuting time, distance from j, geographical and cultural amenities, and many others. Also included are occupational characteristics such as working conditions, relevance to previous human capital investments, coincidence with personal preferences, and flexibility of hours, among many others.

2.1.3 Residuals

To show concretely how preferences for and earnings in a location-occupation combination might affect decisions, consider deviations from the mean of each component in (2.2):

$$w_{i\ell k} - \mathbb{E}\left[w_{i\ell k} \mid x_i, s_i\right] = \eta_{i\ell k} \tag{2.3}$$

$$u_{ij\ell k} - \mathbb{E}\left[u_{ij\ell k} \mid z_i\right] = \varepsilon_{ij\ell k} \tag{2.4}$$

[5] Other exclusion restrictions used in this analysis to distinguish preferences for locations and occupations from earnings include spousal employment status (if married), ages of co-resident children, whether a family member is present (if unmarried), and whether the residence is owned or rented.

[6] The validity of this assumption has been analyzed in previous work by Coate (2013). He notes that the effect of birth location on earnings is ambiguous because of two competing forces: (i) individuals who prefer living close to family may forgo higher earnings available elsewhere in order to stay home; and (ii) individuals staying home may have higher earnings because family networks enable a better job match than could be had elsewhere. He finds that the earnings effect is heterogeneous by education level, where the positive force is stronger for high-school educated workers while the opposite is true for college-educated workers. Given the mixed empirical evidence and the absence of other suitable exclusion restrictions, I maintain this assumption throughout the paper.

then

$$\begin{aligned} V_{ij\ell k} &= v_{j\ell k} + e_{ij\ell k} \\ &= \underbrace{\mathbb{E}\left[w_{i\ell k} \mid x_i, s_i\right] + \mathbb{E}\left[u_{ij\ell k} \mid z_i\right]}_{v_{j\ell k}} + \underbrace{\eta_{i\ell k} + \varepsilon_{ij\ell k}}_{e_{ij\ell k}} \end{aligned} \tag{2.5}$$

where $v_{j\ell k}$ is referred to as either the subutility function (in the selection literature) or the conditional value function (in the dynamic discrete choice literature).

2.1.4 Utility maximization

Individuals maximize utility such that

$$d_{ij\ell k} = 1\left[v_{j\ell k} + e_{ij\ell k} \geq v_{jmn} + e_{ijmn} \;\forall\, (m,n) \neq (\ell,k)\right] \tag{2.6}$$

where $1[A]$ is an indicator variable that takes a value of 1 when condition A is true and 0 otherwise. (2.6) emphasizes that utility depends not only on the location of residence, but also on the deterministic and stochastic elements of utility in *each* location, including the location of birth. Furthermore, earnings are observed only in the location that is selected:

2.1.5 Selection rule

The selection rule is given by

$$w_{i\ell k} \text{ observed} \iff d_{ij\ell k} = 1 \tag{2.7}$$

Specifically, earnings are only observed if all L selection equations in (2.6) are simultaneously satisfied. Thus, individuals observed to reside in ℓ are not a random sample of the population; hence

$$\begin{aligned} \mathbb{E}\left[\eta_{i\ell k} \mid w_{i\ell k} \text{ observed}\right] &= \mathbb{E}\left[\eta_{i\ell k} \mid d_{ij\ell k} = 1\right] \\ &= \mathbb{E}\left[\eta_{i\ell k} \mid e_{ijmn} - e_{ij\ell k} \leq v_{j\ell k} - v_{jmn},\ \forall\, (m,n) \neq (\ell,k)\right] \\ &\neq 0 \end{aligned} \tag{2.8}$$

Dahl refers to $\mathbb{E}\left[\eta_{i\ell k} \mid d_{ij\ell k} = 1\right]$ as the selectivity bias for i. If $\mathbb{E}\left[\eta_{i\ell k} \mid d_{ij\ell k} = 1\right]$ is correlated with x_i or s_i then OLS will returned biased estimates.

Equations (2.1) through (2.7) comprise an extended Roy model of earnings, migration, and occupational choice.

Unfortunately, this extended Roy model is difficult to estimate without making additional assumptions about how the subutility functions affect the selection term (i.e. the conditional

expectation in (2.7)). There are two reasons for this: (i) the number of locations L needs to be sufficiently large in migration models in order to accurately reflect the actual choice set faced by individuals, thus effecting the curse of dimensionality; and (ii) individuals derive utility from both earnings and non-pecuniary aspects of the location, meaning that the researcher is required to account for individual preferences. The problem with the latter reason is that there are a large number of variables that are important factors in the non-pecuniary dimension, but which are unobserved or poorly measured.

In the next section, I explain how I avoid these issues by implementing existing estimation methods (Lee, 1983; Dahl, 2002) which are designed to circumvent parametric estimation of the subutility functions, and which work well on choice sets that are otherwise prohibitively large.

3 Reducing the Dimensionality of the Problem

Estimating the problem described in Section 2 is infeasible without making additional assumptions. The difficulty arises out of the curse of dimensionality due to the large set of locations and occupations in which a person can choose to live and work. In this section, I provide intuition and a brief formal derivation on how to feasibly estimate the aforementioned extended Roy model. The key point is that I follow the strategy developed by Lee (1983) and refined by Dahl (2002) to express the selection in the earnings equation as a function of a small number of observed choice probabilities.

The intuition of this approach is as follows: examining equations (2.6) and (2.7) reveals that the probability of observing an individual's earnings in location ℓ and occupation k is related to the probability that $V_{j\ell k}$ is the maximum of all subutility functions. Thus, the joint distribution between the error term in the earnings equation ($\eta_{i\ell k}$) and the differenced subutility error terms ($e_{j11} - e_{jmn}, \ldots, e_{jLK} - e_{jmn}$) can be reduced from $L \times K$ dimensions to two dimensions: the first dimension is the earnings error and the second is the maximum order statistic of the differenced subutility functions. The key assumption is that this bivariate distribution does not depend on the subutility functions themselves, except through a small number of choice probabilities.[7] This allows the researcher to express the selection correction term in the earnings equation (analogous to the inverse Mills ratio term in the canonical Heckman selection model) as a function of a small number of observed choice probabilities. Without this assumption, the researcher would be required to estimate an $(LK - 1)$-dimensional integral. This becomes quickly infeasible as L grows large, as is the case in the current setting.

To aid the exposition, I now briefly formalize the above intuition. Readers interested in a more detailed derivation should consult Dahl (2002) and Lee (1983).

[7] Dahl (2002) refers to this assumption as the "index sufficiency assumption," which I discuss below in more detail.

First consider a reformulation of (2.6) and (2.7):

$$\begin{aligned} w_{i\ell k} \text{ observed} &\iff v_{j\ell k} + e_{ij\ell k} \geq v_{jmn} + e_{ijmn} \;\forall\, (m,n) \neq (\ell,k) \\ &\iff \left(v_{j11} - v_{j\ell k} + e_{ij11} - e_{ij\ell k}, \ldots, v_{jLK} - v_{j\ell k} + e_{ijLK} - e_{ij\ell k}\right)' \leq 0 \\ &\iff \max_{m,n} \left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \leq 0 \end{aligned} \tag{3.1}$$

Now consider the joint cumulative distribution $F_{j\ell k}$ of the earnings equation error term in (2.1) and the selection rule error terms in (2.6), respectively evaluated at a constant r and the corresponding difference in subutility functions:

$$\begin{aligned} F_{j\ell k}\left(r, v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) &= \Pr\left(\eta_{i\ell k} < r, e_{ij11} - e_{ij\ell k} < v_{j11} - v_{j\ell k}, \right. \\ &\qquad \left. \ldots, e_{ijLK} - e_{ij\ell k} < v_{jLK} - v_{j\ell k}\right) \\ &= \Pr\left(\eta_{i\ell k} < r, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \leq 0\right) \\ &= G_{j\ell k}(r, 0) \end{aligned} \tag{3.2}$$

Re-expressing this in terms of probability density functions, we have the following one-to-one mapping between the LK-dimensional density $f_{j\ell k}$ and the two-dimensional density $g_{j\ell k}$. This mapping is made possible by implementing maximum order statistics (see Lee, 1983):

$$\begin{aligned} &f_{j\ell k}\left(\eta_{i\ell k}, e_{ij11} - e_{ij\ell k}, \ldots, e_{ijLK} - e_{ij\ell k}\right) \\ &= g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \end{aligned} \tag{3.3}$$

where the expression for $g_{j\ell k}$ in (3.3) is written as being conditional on the differences in the subutility functions in order to emphasize this dependence.

Rewriting the earnings equation in (2.1) to correct for selection would yield

$$w_{i\ell k} = \gamma_{0\ell} + x_i\gamma_{1\ell} + s_i\gamma_{2\ell k} + \sum_{j=1}^{L}\sum_{k=1}^{K} d_{ij\ell k}\psi_{j\ell k}\left(v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) + \eta_{i\ell k}, \tag{3.4}$$

where $d_{ij\ell k}$ is as defined in (2.6) and $\psi_{j\ell k}(\cdot) = \mathbb{E}[\eta_{i\ell k} \mid \cdot]$. Dahl notes that (3.4) is called a partially-linear, multiple-index model because it combines a linear model with a set of non-linear control functions $\psi_{j\ell k}$ of the multiple indices $v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}$. Because $\psi(\cdot)$ still depends on the subutility associated with each choice alternative, (3.4) suffers from the curse of dimensionality.

In order to reduce the dimensionality of the selection correction term $\psi(\cdot)$ in (3.4), Dahl

proposes an *index sufficiency assumption* as follows:

$$g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid v_{j11} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \tag{3.5}$$
$$= g_{j\ell k}\left(\eta_{i\ell k}, \max_{m,n}\left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right) \mid p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right)$$

where $p_{ij\ell k}$ is the probability that i moves from location j to location ℓ and works in occupation k, p_{ijmn} is the corresponding probability of moving from j to m and working in n, and $p_{ijm'n'}$ is the corresponding probability of moving from j to m' and working in n'. Note that $p_{ij\ell k}$ corresponds to the individual's observed first-best choice. I discuss later how the other probabilities are constructed.[8] The implicit assumption in (3.5) is that the probabilities $p_{ij\ell k}$, p_{ijmn}, and $p_{ijm'n'}$ contain all of the information about how the index of subutility functions influences the joint distribution of the earnings error term and the maximum of the subutility errors.

Applying the assumption in (3.5) to the earnings equation gives the following corrected earnings equations that account for selective migration and occupational choice, and that are feasibly estimated:

$$w_{i\ell k} = \gamma_{0\ell} + x_i\gamma_{1\ell} + s_i\gamma_{2\ell k} + \sum_{j=1}^{L}\sum_{k=1}^{K} d_{ij\ell k}\lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right) + \omega_{i\ell k}, \tag{3.6}$$

The implication of the assumption in (3.5) is that $\mathbb{E}\left[\omega_{i\ell k} \mid x_i, s_i, p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}, d_{ij\ell k} = 1\right] = 0$, meaning that the selection problem has been resolved. Note also that the index sufficiency assumption reduces the dimensionality of the selection correction functions from LK, LK-dimensional control functions to LK bivariate control functions.

It is important to recognize the restrictions that the index sufficiency assumption levies. Dahl (2002) discusses at length the types of models that satisfy index sufficiency and concludes that the assumption is generally innocuous. He also provides proofs that $\psi_{j\ell k}(\cdot)$ is equal to $\lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right)$. Additionally, he provides evidence from Monte Carlo simulations that $\lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right)$ appropriately corrects for selection. Such proofs are beyond the scope of this article.

In Section 5, I discuss details of the estimation of equation (3.6) including how to estimate the probabilities of interest, and how to estimate the unknown correction functions $\lambda_{j\ell k}$, including additional assumptions made to reduce the number of control functions entering (3.6).

[8] Dahl discusses which of many potential probabilities would best serve as additional probabilities in (3.5). He settles on defining the set of probabilities as $p_{ij\ell}$ and p_{ijj} (if $j \neq \ell$). Note that he does not model occupational choice.

4 Data and Descriptive Analysis

I now discuss in detail the process used to construct the data that my estimation procedure is based on. I also present a descriptive analysis of the data trends which, when compared with the model estimates, will be used to quantify the amount of selection in migration and occupation decisions.

4.1 Data

I use data from the American Community Survey (ACS) as compiled by Ruggles et al. (2015) over the years 2010-2014. The ACS is an annual stratified random sample of 1% of US households produced by the US Census Bureau. Sampled households respond to the survey either on paper or via the internet, and non-responding households receive a follow-up telephone call or visit by a Census employee.

The ACS collects detailed data for each adult household member on income, employment, education, demographic characteristics, and health. It also collects information about the household, such as household and family structure and housing unit characteristics. In this analysis, I focus on the following variables: location of birth, location of residence, residence ownership status, demographic characteristics (e.g. age, gender, race, ethnicity, household composition), education level (including college major), occupation, and earnings.[9]

The analysis sample consists of all native-born individuals between the ages of 22 and 54 with exactly a bachelor's degree, and who have observed earnings within a reasonable range, who have observed college major, who are not in school, do not live in group quarters, and who do not have imputed values for any of the variables of interest. This corresponds to a 5% sample of the relevant US population. The estimation sample of the data comprises 593,848 individuals. Details on the number of observations deleted with each criterion are listed in Table A1.

4.1.1 Data Construction

This section details the steps followed in creating each of the key variables of interest.

Education level I define education level as taking one of M different values, where M is the number of distinct college major categories. In theory, the ACS records hundreds of distinct college major fields. However, in order to focus the analysis and to maintain statistical power, I aggregate such that $M = 5$. The set of aggregated majors is: education, social sciences, business, STEM, and all others. A detailed mapping of the 51 Department of Education major fields to these five aggregated fields is provided in Table A2. Notably, the business field includes economics majors and the STEM field includes pre-med majors.

[9] Information on college major began to be collected in 2009. I focus on the years 2010-2014 in order to maximize sample size while avoiding the most severe part of the Great Recession.

Occupation I define occupation as having two values: "related" or "unrelated" (i.e. $K = 2$). I define an occupation as related to a major if it is reported to have a 2% or larger share of all 3-digit occupation codes within a detailed definition of major (i.e. the 51 Department of Education codes).[10] The set of occupations that are related to an aggregated major category is then the union of the set of related occupations for each of the detailed majors corresponding to the aggregate.

A list of related occupations for each of the 5 college major categories is listed in Table A3. Broadly speaking, the list of related occupations makes sense, and coincides with other papers in the literature.[11] Importantly, the definition of relatedness explained here does not preclude the same occupation from being related to two different majors. This distinction allows for the occupation relatedness definition to match what is observed in the data.

To further illustrate my definition of occupation relatedness, I discuss four different extremes observed from Table A3. First, engineering occupations are not considered to be related to any major except STEM. Second, miscellaneous administrators are considered to be related to every major. Third, lower-level service jobs in food services, tourism, and administrative support tend to only be related to other majors, reflecting the occupations that aspiring performing artists tend to work in. Finally, accountants and auditors are related to business majors, other majors, and STEM majors. Based on these illustrative examples, my definition of occupation relatedness seems to be reasonable.[12]

Race and ethnicity I construct a measure of race and ethnicity by first assigning anyone of Hispanic origin to be Hispanic, and then assigning race based on whether the reported race is white, black, or other. Mixed-race individuals are classified as other.

Earnings and employment Earnings are measured as the individual's annual wage and salary income, expressed in constant 2010 dollars. I drop any nominal earnings measurements greater than \$600,000 or less than \$20,000. I classify a person as employed if they reported being employed at the time of the survey. I also create a variable indicating if the individual's spouse is employed.

[10] This is similar to the "Top 5" occupation distinction made by Altonji et al. (2016b), but is more flexible in defining relatedness by taking into account the distribution of occupations within a given major.

[11] As an example, Kinsler and Pavan (2015) use a self-reported measure of occupational relatedness and find that there is considerable overlap across majors among workers who report being in the same related occupation. The difference between my definition of relatedness and the self-reported definition in Kinsler and Pavan is that my approach restricts all individuals in an occupation-major category to be either related or unrelated. In contrast, the self-reported definition of relatedness allows for both unrelated and related jobs to be observed in every occupation-major category.

[12] Note that, because I focus on individuals with exactly a bachelor's degree, advanced professional degree occupations such as lawyers, doctors, and professors are excluded from this list.

Work experience I define work experience as potential experience in the usual way: age minus number of years of schooling minus 6.

Birth place I create separate variables indicating in which state the individual was born, and in which state the individual's spouse was born (if applicable).

Marital status and household composition Marital status is self-reported in the survey as one of six categories. I aggregate these categories into three: married (whether or not residing with spouse); divorced or separated; and single or widowed. Number of co-resident children is given in the survey and I distill this information into two dummies: one or more children under the age of 5; and one or more children under the age of 18. Family co-residence status is distilled into one dummy variable indicating whether the individual is in the same household as any relative. The relationship can be blood, or through marriage.

Dwelling characteristics Home ownership status is divided into "owned" or "rented."

4.2 Descriptive Analysis

I now discuss some descriptive evidence from the data that motivate my treatment of location- and occupation-specific college major premiums.

I first present in Table 1 overall summary statistics for the estimation sample. Business and STEM majors have the highest earnings. Education and business majors are the least likely to move away from their state of birth, while other majors and STEM majors are most likely to move. Education majors are also the most likely to work in a related occupation, followed by STEM and Business majors. From a demographic perspective, education, social science, and other majors are disproportionately female, while social science majors disproportionately represent minorities.

Taken together, the results of Table 1 paint a complex, multi-dimensional picture of the labor market outcomes of different college majors. Education majors have the lowest earnings and the lowest moving propensity, but the highest related occupational employment propensity. On the other hand, Science majors second among the five groups in terms of earnings, migration propensity, and related occupational employment propensity. These facts motivate an analysis of location-specific outcomes and location-specific preferences that are allowed to differ by college major.

To analyze location-specific outcomes, I estimate regressions of the form

$$y_{i\ell} = \gamma_{0\ell} + x_i\gamma_{1\ell} + s_i\gamma_{2\ell} + \eta_{i\ell}, \quad \ell = 1, \ldots, L \tag{4.1}$$

where $y_{i\ell}$ is a vector of outcomes and x_i and s_i are defined as in (2.1). I then plot the frequency distribution of the estimated major dummies $\hat{\gamma}_{2\ell}$, where education major is the refer-

ence category. The other covariates in these regressions include a cubic in potential experience, and the following sets of dummies: married, female, race categories, and specific metropolitan areas.[13]

Figure 1 plots these distributions for each major and reveals several interesting findings. First, education majors earn the lowest in all locations except a small handful. Second, social science majors and other majors have similar mean earnings, but social science majors have a higher variance in the cross-location earnings distribution. Similarly, business and STEM majors have similar means, but business majors have a higher variance across locations. There is little overlap between the high-earning business and STEM distributions and the lower-earning social science and other distributions.

I next examine the frequency of migration by college major and destination location. To do so, I estimate (4.1), but where now $y_{i\ell}$ a dummy indicating that the individual has moved away from her state of birth.The frequency distributions of $\hat{\gamma}_{2\ell}$ are displayed in Figure 2. The graphs illustrate that education majors are the least mobile, with the exception of business majors in a small number of locations. STEM and other majors are the most mobile. Overall there is more overlap and less variance across majors in this figure than in Figure 1. This motivates allowing for nonpecuniary factors to influence migration decisions.

The discussion up until now has abstracted from occupational choice. To see how occupational relatedness is concentrated across space, I estimate (4.1)but where now $y_{i\ell}$ a dummy indicating that the individual works in an occupation related to her major. Occupational relatedness is defined above in Section 4.1.1. The results in Figure 3 show that education majors are far more likely to work in a related occupation than any other major in most locations. Business and STEM majors show the highest levels of occupational relatedness, even surpassing education majors for about half of locations. Like the migration distributions in Figure 2, the large amount of overlap indicates that nonpecuniary preferences might be a large determinant of occupational choice. This evidence supports my decision to model location-specific occupational choice.

I now investigate the effect of working in a related occupation on earnings by estimating a descriptive model similar to (4.1), but allowing for location- and occupational-specific parameters.

$$y_{i\ell k} = \gamma_{0\ell} + x_i\gamma_{1\ell} + s_i\gamma_{2\ell k} + \eta_{i\ell k}, \quad \ell = 1, \ldots, L\ , \ k = 0, 1 \tag{4.2}$$

where the primary difference relative to (4.1) is that the major dummies $\gamma_{2\ell}$ are fully interacted with a dummy for occupational relatedness.

I plot in Figure 4 the frequency distributions of the differenced location-specific estimates $\hat{\gamma}_{2\ell 1} - \hat{\gamma}_{2\ell 0}$, which correspond to the within-location premium associated with working in a related occupation. Interestingly, these distributions mimic the raw major premium distribu-

[13]For example, if the location is California, the metropolitan area dummies will indicate residence in Los Angeles, San Francisco, San Diego, San Jose, Riverside, etc.

tions listed in Figure 1: STEM and business majors earn the most from working in a related occupation, followed by other majors, then social science majors, with education majors being the lowest. Interestingly, the premium for working in an occupation related to a STEM major exhibits a slightly bimodal shape across locations, with modes at approximately 12% and 22%. Another interesting finding from Figure 4 is that education majors face a *negative* occupational relatedness premium in roughly half of all locations. This again highlights the role of nonpecuniary preferences in the occupation as well as the location decision. For instance, if an education major has a strong preference for working as a teacher, why would she not move to the location that has the highest premium to such a decision?

Finally, I examine the extent to which the distributions discussed above are simply due to persistent characteristics about the chosen location rather than the chosen college major. For example, is it the case that the location rank of earnings is the same for each major? Does location explain occupational relatedness more than major? To analyze this dimension, I compute the correlation matrix across majors for the four outcomes depicted in Figures 1 through 4. Table 3 presents these correlations. The most striking finding is in panel (c) which shows that occupational relatedness appears to be a local characteristic much more than a major characteristic. Earnings appear to be more locally determined than not. There is weaker evidence that certain locations are "in-migrant" locations.[14] Finally, the premium to working in a related occupation seems to be completely independent of location and thus determined by major. This last piece of evidence motivates modeling the link between migration and occupational choice.

I emphasize that the results of Figures 1 through 4 are contaminated with selection bias. In order to better understand how much of the observed major-specific earnings premiums are due to selectivity in location and occupation decisions, it is necessary to use the framework described in Sections 2 and 3.

4.3 Transition Matrix

The results of the previous section indicate that there is substantial heterogeneity in the observed earnings levels, occupational propensities, and migration behavior of different majors in different states. I now examine the heterogeneity in migration and occupational choice across majors originating and residing in different pairs of states.

Figure 5 displays the migration transition matrix by major for the five largest states. Rows indicate birth location, while columns indicate residence location. The bottom section of each bar corresponds to the related occupation, while the top section corresponds to the unrelated.

Upon examining Figure 5, a number of motivating facts stand out. First, Texas and California appear to be popular destinations for all majors. At the same time, Florida in particular is a popular destination for New Yorkers, and more so for education majors. Interestingly, Floridian

[14] This finding is related to Zabek (2016) who shows that growing locations tend to have a higher fraction of in-migrants, while declining locations tend to have a higher fraction of stayers.

and Texan education majors work in related occupations at a much higher rate than their New York or California counterparts. This is one explanation for why Florida is such an attractive destination for New York education majors.

These results and others in the figure provide additional evidence of the presence of nonpecuniary factors on the decision to migration and choose an occupation.

5 Estimation

In this section, I discuss how to estimate the final equation (3.6) of the model discussed in Sections 2 and 3. The estimation proceeds in two stages. First, I estimate the migration probabilities $\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right)$. Second, I estimate the parameters of equation (3.6), including the unknown correction functions $\lambda_{j\ell k}$.

5.1 Migration probabilities

There are a variety of ways in which one can estimate the migration probabilities. Some alternatives include the conditional logit model, the conditional probit model, or non-parametric estimation techniques.

The conditional logit model is by far the most popular in estimating migration probabilities (a setting where the dimension of the choice set is large) due to its simple closed-form expression for the underlying choice probabilities. The primary drawback of this model is that it suffers from the independence of irrelevant alternatives property.[15]

The conditional probit model (Hausman and Wise, 1978) allows for arbitrary correlations among the choice alternatives, but is unsuitable for settings such as this where the choice set is large. This is because the conditional probit model requires estimation of a $(J-1)$-dimensional integral, where J is the number of alternatives. Using this model would would eliminate the gains afforded by the index sufficiency assumption discussed in Section 3. The conditional probit model also requires the researcher to specify the covariance structure of the alternatives.

Non-parametric estimation has two advantages. First, it does not require the researcher to model location-specific characteristics, of which there are an inordinate number and many of which are poorly measured. Second, it does not require the researcher to specify the dependence structure of the choice alternatives as would be required with the conditional probit model.[16]

The primary drawback to non-parametric estimation is deciding how finely and in which ways to divide the state space. Probabilities that are estimated from cells that are too small will

[15]For tractability reasons, dynamic migration models such as Kennan and Walker (2011) and Ransom (2016) assume that migration probabilities take a conditional logit form. Davies et al. (2001) assume this form in a static setting. Monras (2015) argues that a nested logit is more appropriate for characterizing migration decisions.

[16]Hausman and Wise (1978) note that the conditional probit model produces inconsistent estimates of the choice probabilities if dependence among the alternatives is incorrectly assumed. Likewise, the conditional logit model produces inconsistent estimates if there is in fact any dependence among the alternatives.

introduce a large amount of error into the estimation. On the other hand, failure to create enough cells will result in probabilities that do not accurately represent the data.

5.1.1 Non-parametric estimation using machine learning

I estimate the location and occupational choice probabilities non-parametrically using a method from the machine learning literature called conditional inference recursive partitioning, developed by Hothorn et al. (2006) and implemented in the R programming language by Hothorn and Zeileis (2015).

The algorithm is designed to overcome the drawbacks associated with non-parametric estimation. The main advantage is that it prevents the researcher from being required to make *ad hoc* assumptions about how the state space should be divided when creating probability bins. It also has the advantage of automatically merging together sparse bins such that the algorithm does not return any empty bins or any bins of excessively small size. I detail the conditional inference tree algorithm in the following subsection.

Generally speaking, machine learning methods combine estimation with model selection to enhance out-of-sample prediction. In the current setting, the conditional inference recursive partitioning algorithm selects which variables and which categories of the variables matter most in predicting migration and occupations. For other settings where the set of covariates is larger than the sample size, model selection methods automatically choose which covariates should be included such that standard rank and order conditions for identification are satisfied.[17] Varian (2014) provides an overview of basic machine learning algorithms and suggests ways in which they can be used to improve existing research methods in economics. Other examples of machine learning applications in economics include Athey and Imbens (2015), Gentzkow et al. (2015), and Belloni et al. (2011).[18]

5.1.2 Conditional inference recursive partitioning algorithm

The conditional inference recursive partitioning algorithm is a classification tree algorithm designed to non-parametrically predict a dependent variable from a set of covariates. The algorithm takes as inputs the dependent variable and the covariates, and returns as outputs combinations of the covariates that form clusters (nodes of the tree) or cells. Using an internal stopping criterion based on hypothesis testing, it optimally trades off bias (creating too few clusters and, as a result, poorly fitting the estimation data) and variance (creating too many clusters and, as a

[17] This setting applies to Bajari et al. (2015) who show how a variety of machine learning methods can be used in demand estimation to evaluate advertising effectiveness.

[18] Athey and Imbens (2015) show how machine learning methods can be used to estimate heterogeneous treatment effects. Gentzkow et al. (2015) illustrate how to use model selection to estimate polarization in high-dimensional textual data. Belloni et al. (2011) develop methods for using model selection in instrumental variables models when the number of instruments is larger than the sample size.

result, poorly fitting out of sample) such that out-of-sample prediction is maximized.[19] The algorithm works for both continuous and categorical variables on both sides of the equation.[20] The current application contains a categorical dependent variable and covariates that are primarily categorical, but some of which are continuous.

Below, I detail the algorithm, which recursively iterates on the following two steps:

1. *Selection.* The algorithm begins by testing whether the dependent variable is independent of the covariates (i.e. testing whether the distribution of the dependent variable Y is different from the conditional distribution $Y|X_j$ for all covariates). If any member of this set of conditional distributions is significantly different from the unconditional distribution, then the algorithm selects the covariate with the strongest association with Y as measured by a p-value.

2. *Splitting.* Once a covariate has been selected, the algorithm optimally splits it. This is done in a similar fashion as the selection, only the algorithm at this phase selects among different *subsets* of the specified covariate. The optimal split is the one that creates the most distinct pair of distributions of the dependent variable, as measured by a p-value. There are other criteria involved in determining if a candidate split is carried out; namely how large the resultant cluster will be. Clusters that are too small will predict poorly out-of-sample and are skipped accordingly.

The algorithm then iterates on these two steps until at least one of the following criteria is met:[21]

- No additional covariates can be selected because they fail to reject the null hypothesis of independence.
- Any further splits of the already-selected covariates would fail to reject the null hypothesis of equality in the dependent variable across the split
- Any further splits would result in clusters with too few observations (i.e. unsuitable for out-of-sample prediction)
- The candidate cluster already perfectly predicts the dependent variable

[19] Hothorn et al. (2006) emphasize that the internal stopping criterion acts similarly to pruning or cross-validation methods that are commonly used in other machine learning settings to penalize complexity.

[20] In the case of a continuous dependent variable, the algorithm minimizes the sum of squared errors within each cluster to find the optimal cluster division. In the case of a continuous covariate, the algorithm creates bins by choosing cut points. Additionally, the algorithm can also be used in survival analysis.

[21] There are a few tuning parameters of the algorithm that the researcher can adjust. One is the p-value that determines splitting, another is the smallest number of observations allowed in a cluster, and a third is the smallest number of observations allowed in a candidate node split (i.e. the minimum number of observations required in each resulting subset of the split). I choose 5% for the p-value parameter, 50 observations for the minimum cluster size, and 50 observations for the minimum candidate node split size. These were chosen via cross-validation, but in practice the predictive accuracy of the tree algorithm was not sensitive to these tuning parameters.

- No further splits are possible because the candidate cluster is composed of a single combination of all independent variables

As an example of what the output of this algorithm looks like, I include Figure 6, which depicts a simple example of the output from a fictitious migration dataset. Individuals are characterized only by their level of work experience and can choose to live in 3 locations: New York, Texas, or elsewhere. The algorithm shows that experience is the strongest predictor of location choice, and that the most distinct difference occurs when splitting at three, followed by an additional split that occurs at eight. The algorithm shows that New York is entirely composed of individuals with less than four years of work experience, that Texas is composed nearly perfectly of individuals with experience levels between four and eight years, and that workers with nine or more years of experience almost certainly live elsewhere. In the actual estimation, each node will be composed of 102 categories (rather than three), and each tree will typically have many more than three terminal nodes.

5.1.3 Implementation of the non-parametric estimation algorithm

I now discuss in detail the estimation of the migration probabilities and which variables are used to predict migration and occupational choice. Following Dahl (2002), I use cell decision probabilities, where the cells are computed from the recursive partitioning algorithm detailed above. The implicit assumption with this approach is that observably similar people face similar unobserved earnings and preference shocks. Importantly, this implies that the researcher need not model the characteristics of the alternatives, only the characteristics of the individuals.

Formally, the cell migration probability for all individuals, all origin locations j, and all destination locations ℓ and occupations k is

$$\begin{aligned} p_{ij\ell k} &= \Pr\left(d_{ij\ell k} = 1 \,|\, v_{j1k} - v_{j\ell k}, \ldots, v_{jLK} - v_{j\ell k}\right) \\ &= \Pr\left(d_{ij\ell k} = 1 \,|\, \text{cell}\right) \end{aligned} \tag{5.1}$$

The conditional inference tree algorithm assigns cells based on the following characteristics: whether the individual was born in any of the following locations: the state of residence, an adjacent state, within the same Census division, or the same Census region; college major; age; race; gender; marital status; whether the current residence is owned or rented; whether or not the individual is living with a family member or relative; whether or not the individual's spouse is working (if married); the presence of children ages 0-4 and ages 5-18; and the popularity of related occupations in the state of birth for the given demographic cell (the exclusion restriction governing the occupational relatedness choice). I estimate the cell probabilities using the so-called "one-vs-all" classification method: for each residence location and occupation, I compute the probability of belonging to the choice alternative under consideration.

5.1.4 Tree algorithm performance relative to more commonly used methods

A valid question regarding the conditional inference tree algorithm is how it compares with a non-parametric bin estimator or to a simple logit estimator. To assess the performance of each of the estimators, I estimate the first-best choice probabilities for each algorithm using the 2010-2014 ACS sample discussed previously. I then test the out-of-sample predictive performance of each algorithm using a holdout sample of the 2010-2014 ACS. The results from this exercise are detailed in Table A4. Each of the three classification algorithms performs similarly in terms of raw predictive accuracy as well as penalized predictive accuracy. However, the tree algorithm provides a much greater level of variation for the related occupation exclusion restriction, which I detail later. This highlights the usefulness of the tree algorithm in allowing different divisions of the state space for different observations. The definitions of each of these accuracy metrics are detailed in Table A4. It is interesting to note that the multinomial logit slightly outperforms both of the nonparametric methods. While theoretically unappealing because of the IIA property, the logit performs well from a pure predictive standpoint. This suggests that other logit-based machine learning algorithms such as neural networks could possess even higher levels of predictive accuracy of migration.

5.2 Correction functions

I now describe how to feasibly estimate the unknown selection correction functions in (3.6). As written, this equation contains LK bivariate correction functions for each location ℓ and occupation k. To further simplify this, I follow Dahl and make the assumption that the selection correction functions are the same for movers, regardless of the location of origin. In formal terms, this assumption imposes that the correction term in (3.6) be rewritten as

$$\begin{aligned} \lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right) &= \lambda_{\ell k}\left(p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}\right),\ j \neq \ell,\ k \in \{0,1\} \\ \lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}\right) &= \lambda_{jk}\left(p_{ijjk}, p_{ijmn}\right),\ j = \ell,\ k \in \{0,1\} \end{aligned} \tag{5.2}$$

A simplifying assumption akin to the one made in (5.2) is required in order to maintain identification power. If migration were a more common occurrence in the data, so that the underlying cells were more densely populated, it would be possible to estimate separate correction functions for different origin locations.

I now discuss my choice for the probabilities $p_{ij\ell k}, p_{ijmn}, p_{ijm'n'}$. I assign as $p_{ij\ell k}$ the first-best choice probability, which is readily observable in the data. For p_{ijmn}, I use the probability that individual i would stay in the first-best location, but work in the non-chosen occupation. This is simply $p_{ij\ell,k'}$, where k' denotes the non-chosen occupation. Finally, for $p_{ijm'n'}$, I use the probability that individual i would stay in her birth location, summing over both occupation probabilities within that location. This is analogous to the retention probability used in Dahl (2002).

To estimate the unknown correction functions $\lambda_{\ell k}$, I use a flexible polynomial function of the probabilities as discussed in Dahl (2002). For each birth location, there are 28 additional variables in the regression. Specifically, I interact the following polynomial of probabilities with each of the four dummies for the observed migration-occupation category (stayer or mover crossed with related or unrelated): $p_{ij\ell}$ and its square, p_{ijmn} and its square; $p_{ijm'n'}$ and its square; and the three pairwise linear interactions between $p_{ij\ell k}$, p_{ijmn}, and $p_{ijm'n'}$. The resulting equation is of the same form as (3.6), where $L = 2$ (i.e. stay or move) and $K = 2$, and where $\lambda_{\ell k}$ is approximated by the polynomial function just described.[22]

5.2.1 Exclusion restrictions

In order to distinguish between preferences and earnings (and thus identify the selection correction functions), there need to be covariates which affect the decision probabilities, but which do not appear in the earnings equation. These covariates are as follows: state of birth, co-residence with a family member, spouse's work status, spouse's birth place, presence of children aged 0-4 or 5-18, and home ownership status. Following Dahl, I allow birth state and demographic variables to affect migration. For occupational choice, I calculate the the share of workers in the individual's college major and birth state who are working in a related occupation, adjusted for demographic characteristics. This exclusion restriction is similar in spirit to that implemented by Kinsler and Pavan (2015). In this sense, the labor market characteristics of the individual's birth location can be thought of as a pre-market factor that contributes to her occupational choice (Speer, 2016).

5.3 Earnings equation

The parameters of the earnings equation parameters in (2.1) are estimated by OLS after making use of the index sufficiency assumption in (3.5) and the dimensionality reduction assumptions discussed in the previous section.

The standard errors of the parameters associated with the selection functions must be adjusted to account for two elements of the estimation: (i) the selection probabilities are not i.i.d. across individuals because of the cell assumption in (5.1); and (ii) the estimation of the cell probabilities induces estimation error into the coefficients because the true probabilities are not observed. Clustering the standard errors by decision cell (rather than by individual) resolves (i). To resolve (ii), an additional formula that resembles the outer product of the gradients is required, inserted into the standard clustering formula:

$$V = (X'X)^{-1} \left\{ \sum_c u_c' u_c \right\} (X'X)^{-1} \tag{5.3}$$

[22] Note that, for stayers, the retention probability drops out of the polynomial and the selection terms have only five polynomial components instead of nine.

where X is the matrix of earnings equation covariates (including the correction function terms) and u_c is a term that accounts for the fact that the probabilities are estimated at the cell level instead of the individual level:

$$u_c = \sum_{i \in c} e_i x_i$$

where e_i is the OLS residual for individual i, and x_i is the covariate vector for individual i, including the 28 selection probability terms.

6 Empirical Results

In this section, I discuss the results of the estimation procedure described in the previous section. I first present results on the estimation of the decision probabilities. I follow this by discussing the estimates of the returns to majors and occupational relatedness.

6.1 Choice probabilities

The estimated choice probabilities are reported by mover status and chosen occupation in Tables 4 and 5. Each table lists components of the decision probability distribution conditional on the listed education level and the observed migration path (i.e. stay or move) and occupational choice (i.e. related or not). The tables also report the number of individuals in each migration-occupation-education classification and the number of different cells contributing to each classification.

An important aspect of Tables 4 and 5 is the relationship between variation in the decision probabilities and identification of the returns to education and occupational relatedness. Specifically, separating the effect of earnings from preferences requires that the decision probabilities across majors within a migration-occupation bin be overlapping. Intuitively, the returns to major can be calculated by comparing individuals in two different majors who have the same selection bias. A similar argument can be used to identify the returns to occupation relatedness within an education category. In this case, identification requires some amount of overlapping between the probability distribution in panels (a) and (b) of each of the tables. Examination of, e.g., panel (a) of Table 3 with panel (a) of Table 4 reveals that there is plenty of overlap in the probability distributions for each major. The same holds true for, e.g., panel (a) of Table 3 and panel (b) of Table 3 which is used to identify the returns to working in a related occupation.

6.2 Earnings

I now discuss and compare the estimates of the earnings equation with and without the selection correction. These results represent the returns to college majors and working in a related

occupation, which may be heterogeneous across majors and across space.

6.2.1 Estimates for specific states

Table 6 lists the full estimates of equation (3.6) with the implemented simplifications discussed in Section 5.2. While I estimate 51 equations, I present detailed results for only the five most populous states: California, Texas, Florida, New York, and Illinois.

The main takeaway from this table is that the selection bias associated with the endogenous choices of where to live and in which occupation to work is quite heterogeneous across chosen occupation and across space. In some states, the selection bias is minimal, while in others, it is quite large. Furthermore, the direction of the bias varies across states, and it also sometimes varies across chosen occupation within the same state.

For example, in unrelated occupations in California and Texas, the OLS estimates of the returns to majors are downward biased in both of the chosen occupations. However, in Florida, Illinois, and New York, the direction of the bias differs based on the chosen occupation. The implication of this heterogeneity is that the corrected return to working in a related occupation will have a different sign across locations.

Of the five largest states, the selection magnitude in Texas is the largest, particularly for majors in the related occupation. In general, the OLS estimates of the return to majors who are working in related occupations are downward biased, with the exception of Florida.

6.2.2 Estimates for all states

I now present in Tables A5 through A8 specific returns to STEM and business majors working in each occupation for all 51 locations. A common theme from each of these tables is that the sign of the selection bias is quite variable across locations. Moreover, there does not appear to be any systematic variation in which states have significantly different corrected returns. The lone exception is Colorado, which exhibits upward bias in OLS estimates of the returns to both STEM and business majors, regardless of occupation relatedness.

I now discuss the returns to working in a related occupation for each of the majors. Table 7 reports moments of the distribution of the return to working in a related occupation, defined as $\hat{\gamma}_{2\ell 1} - \hat{\gamma}_{2\ell 0}$ from equation (3.6). What is interesting is that the average of the distribution of related occupation returns is quite similar for each of the majors, with and without selection correction. This is in spite of the selection correction terms entering in significantly to the earnings equation in almost all states. This finding is consistent with Dahl (2002), who finds that the college wage premium does not narrow after correcting for selection.

6.2.3 Selection bias and the returns to related occupation

To examine the effect of selection on the location-specific returns to working in a related occupation, I present three final results, which respectively plot the difference between the corrected

and uncorrected returns to related occupation and characterize the distribution of these returns across locations.

Figure 7 plots the uncorrected and corrected returns to related occupation for each of the 51 states and five majors, relative to a 45-degree line. For each of the majors, there appear to be more dots above the line, indicating that, on average, OLS is downward biased. However, none of the majors exhibit any overwhelming direction of bias.

Figure 8 plots the corrected and uncorrected distributions of the returns to related occupation for each major. The corrected returns are represented by shaded bars, while the uncorrected returns are transparent bars. For each of the majors, the corrected distribution appears to have both a higher mean and a higher variance.

Finally, I examine which specific components of the return to related occupation contribute to the general downward bias of OLS. Table 8 reports the 10th, 50th, and 90th percentiles of the percent change between the uncorrected and corrected returns to major. The first three columns are for those who work in an unrelated occupation. The second set of columns are for those who work in a related occupation. The final three columns are the difference between the first two sets of columns, and represent the returns to working in a related occupation. Table 8 displays a wide range of heterogeneity in the percentage change in returns to major when correcting for selection. At the median, OLS estimates of the returns to major in an unrelated occupation tend to be upward biased, while the opposite is true for majors in a related occupation. Together, this implies that the corrected returns are downward biased. The magnitude of bias is largest for education and social sciences majors and lowest for STEM and business majors. For the median location, correcting for selective migration and occupational choice increases the returns to working in a related occupation by anywhere from 10% to 30%.

7 Conclusion

This paper examines the extent to which selection into residence location and occupation biases the observed earnings differences across college majors. To analyze this question, I develop and estimate an extended Roy model where individuals have preferences for both earnings and non-pecuniary aspects of given location-occupation pairs.

To estimate the model, I implement the framework of Dahl (2002) and Lee (1983) which allow for feasible estimation of the extended Roy model by expressing the selection in terms of a small number of observed choice probabilities. I estimate the model using data from the American Community Survey from years 2010-2014. I also illustrate the advantages of using machine learning methods to non-parametrically estimate the selection probabilities. The primary advantage of this is in combining model selection and estimation.

I find that the returns to college major are downward biased by selective migration and occupational choice, but that the magnitude of the bias is small. This suggests that nonpecuniary preferences associated with location and occupation are important determinants of earnings

differences across majors.

References

Altonji, Joseph G., Peter Arcidiacono, and Arnaud Maurel. 2016a. The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the economics of education*, vol. 5, eds. Eric Hanushek, Stephen Machin, and Ludger Wößmann. North-Holland: Elsevier Science, 305–396.

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2016b. Cashier or consultant? Entry labor market conditions, field of study, and career success. *Journal of Labor Economics* 34, no. S1:S361–S401.

Arcidiacono, Peter. 2004. Ability sorting and the returns to college major. *Journal of Econometrics* 121, no. 1:343–375.

Athey, Susan and Guido W. Imbens. 2015. Machine learning methods for estimating heterogeneous causal effects. Working paper, Stanford University.

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2015. Demand estimation with machine learning and model combination. Working Paper 20955, National Bureau of Economic Research.

Bayer, Patrick, Shakeeb Khan, and Christopher Timmins. 2011. Nonparametric identification and estimation in a Roy model with common nonpecuniary returns. *Journal of Business & Economic Statistics* 29, no. 2:201–215.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. Lasso methods for gaussian instrumental variables models. Working paper, Duke Fuqua School of Business, Massachusetts Institute of Technology, and Chicago Booth School of Business.

Coate, Patrick. 2013. Parental influence on labor market outcomes and location decisions of young workers. Working paper, Duke Univeristy.

Dahl, Gordon B. 2002. Mobility and the return to education: Testing a roy model with multiple markets. *Econometrica* 70, no. 6:2367–2420.

Davies, Paul S., Michael J. Greenwood, and Haizheng Li. 2001. A conditional logit approach to U.S. state-to-state migration. *Journal of Regional Science* 41, no. 2:337–360.

Diamond, Rebecca. 2016. The determinants and welfare implications of US workers' diverging location choices by skill: 1980-2000. *American Economic Review* 106, no. 3:479–524.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2015. Measuring polarization in high-dimensional data: Method and application to congressional speech. Working paper, Stanford University, Brown University, and Chicago Booth School of Business.

Hausman, Jerry A. and David A. Wise. 1978. A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica* 46, no. 2:403–426.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47, no. 1:153–161.

Heckman, James J. and Bo E. Honoré. 1990. The empirical content of the Roy model. *Econometrica* 58, no. 5:1121–1149.

Heckman, James J. and Christopher Taber. 2008. Roy model. In *The new palgrave dictionary of economics*, eds. Steven N. Durlauf and Lawrence E. Blume. Palgrave Macmillan, 2 ed., 1–9.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15, no. 3:651–674.

Hothorn, Torsten and Achim Zeileis. 2015. partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research* 16, no. 12:3905–3909.

Kennan, John and James R. Walker. 2011. The effect of expected income on individual migration decisions. *Econometrica* 79, no. 1:211–251.

Kinsler, Josh and Ronni Pavan. 2015. The specificity of general human capital: Evidence from college major choice. *Journal of Labor Economics* 33, no. 4:933–972.

Lee, Lung-Fei. 1983. Generalized econometric models with selectivity. *Econometrica* 51, no. 2:507–512.

Monras, Joan. 2015. Economic shocks and internal migration. Discussion Paper 8840, IZA.

Moretti, Enrico. 2012. *The new geography of jobs*. New York: Houghton Mifflin Harcourt.

Ransom, Michael R and Aaron Phipps. 2016. The changing occupational distribution by college major. Working paper, Brigham Young University.

Ransom, Tyler. 2016. The effect of business cycle fluctuations on migration decisions. Working paper, Duke University.

Roy, A.D. 1951. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, no. 2:135–146.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek. 2015. *Integrated public use microdata series: Version 6.0 [machine-readable database]*. Minneapolis: University of Minnesota.

Speer, Jamin D. 2016. Pre-market skills, occupational choice, and career progression. *Journal of Human Resources* Forthcoming.

Varian, Hal R. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28, no. 2:3–28.

Winters, John V. Forthcoming. Do earnings by college major affect college graduate migration? *Annals of Regional Science* .

Zabek, Mike. 2016. Population growth, decline, and shocks to local labor markets. Working paper, University of Michigan.

Figures and Tables

Table 1: Sample means of outcome and demographic variables, by college major

	Education	Soc Sci	Other	Business	STEM	Overall
Log Earnings	10.17	10.32	10.36	10.59	10.57	10.47
Lives outside birth state	35.63	43.41	46.04	41.66	44.92	43.29
Works in related occ.	67.93	45.85	50.31	60.09	62.12	57.36
Female	74.12	59.18	49.32	42.76	41.42	48.39
White	86.53	79.45	83.59	83.74	84.75	83.7
Black	5.82	9.24	6.57	7.11	6.05	6.84
Hispanic	5.36	6.76	5.95	5.02	4.59	5.35
Other race	2.29	4.55	3.89	4.13	4.62	4.11
Frequency	8.11	12.04	22.7	29.02	28.13	100
N	50,934	71,002	132,033	170,068	169,811	593,848

Notes: All variables except for log earnings are expressed in percentage points. Sample weights are included in the computation.

Source: Author's calculations from American Community Survey, 2010-2014.

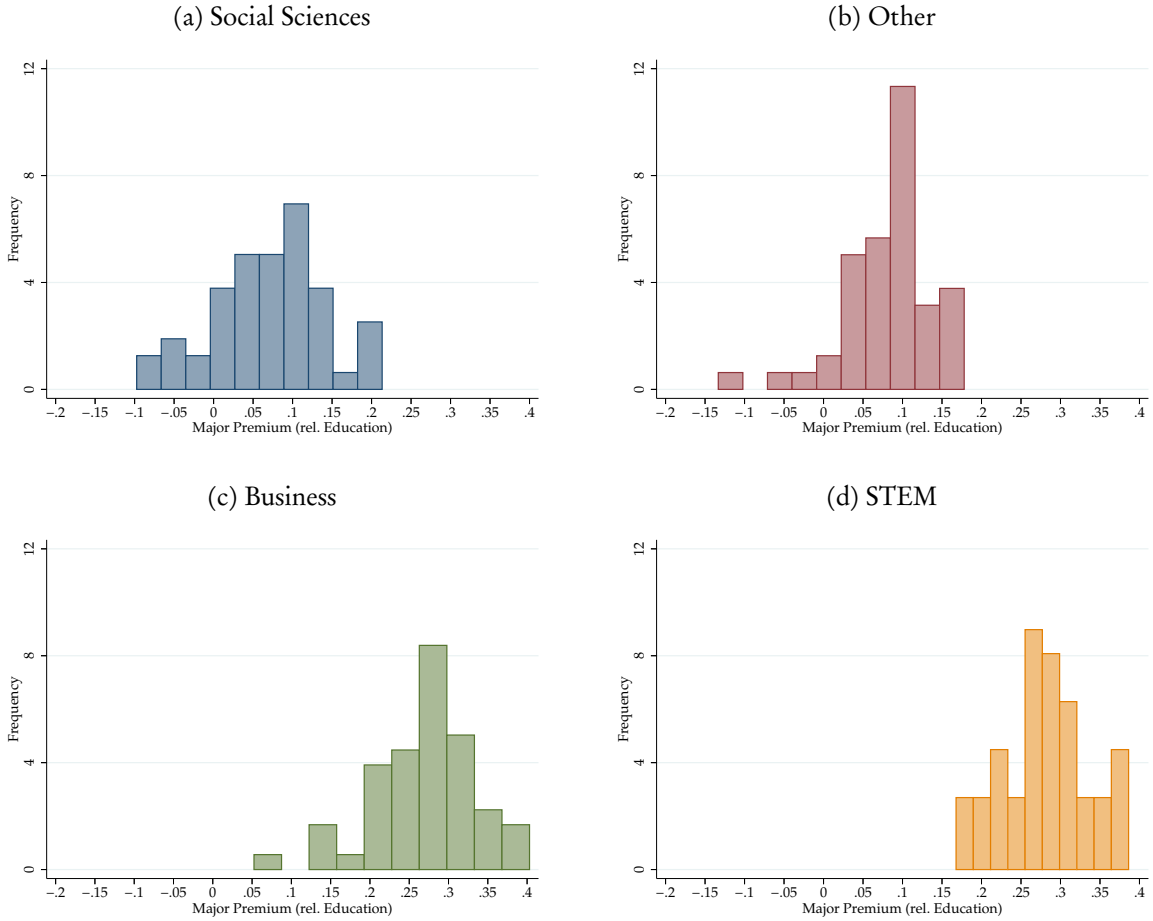
Table 2: Sample means of outcome and demographic variables, by college major (Males only)

	Education	Soc Sci	Other	Business	STEM	Overall
Log Earnings	10.32	10.48	10.47	10.72	10.70	10.62
Lives outside birth state	37.05	46.58	46.83	42.81	47.29	45.26
Works in related occ.	56.19	43.30	52.09	61.77	61.49	57.53
White	85.20	82.19	84.97	86.43	86.16	85.57
Black	7.27	7.64	6.07	5.35	4.61	5.57
Hispanic	4.88	6.06	5.56	4.36	4.54	4.87
Other race	2.64	4.11	3.40	3.86	4.69	4.00
Frequency	4.06	9.53	22.29	32.19	31.93	100
N	13,046	28,581	66,706	96,379	99,583	304,295

Notes: All variables except for log earnings are expressed in percentage points. Sample weights are included in the computation.

Source: Author's calculations from American Community Survey, 2010-2014.

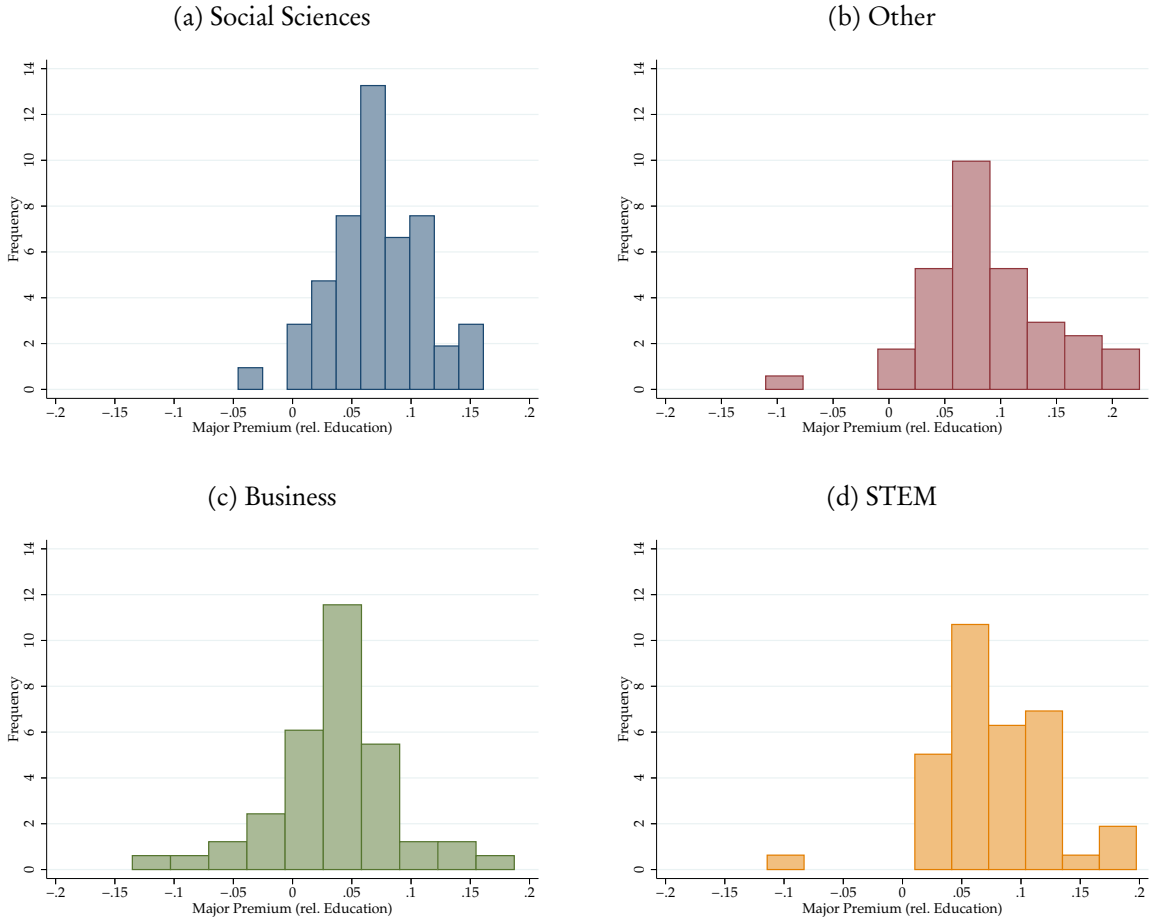
Figure 1: Major-specific earnings distributions across locations



Notes: Above are histograms of the coefficient on major dummies in a log earnings regression conditional on residing in a specific U.S. State. Additional controls in the regression include a cubic in potential experience, gender and race dummies, CBSA dummies, and a married dummy. Census population weights are used in the calculations.

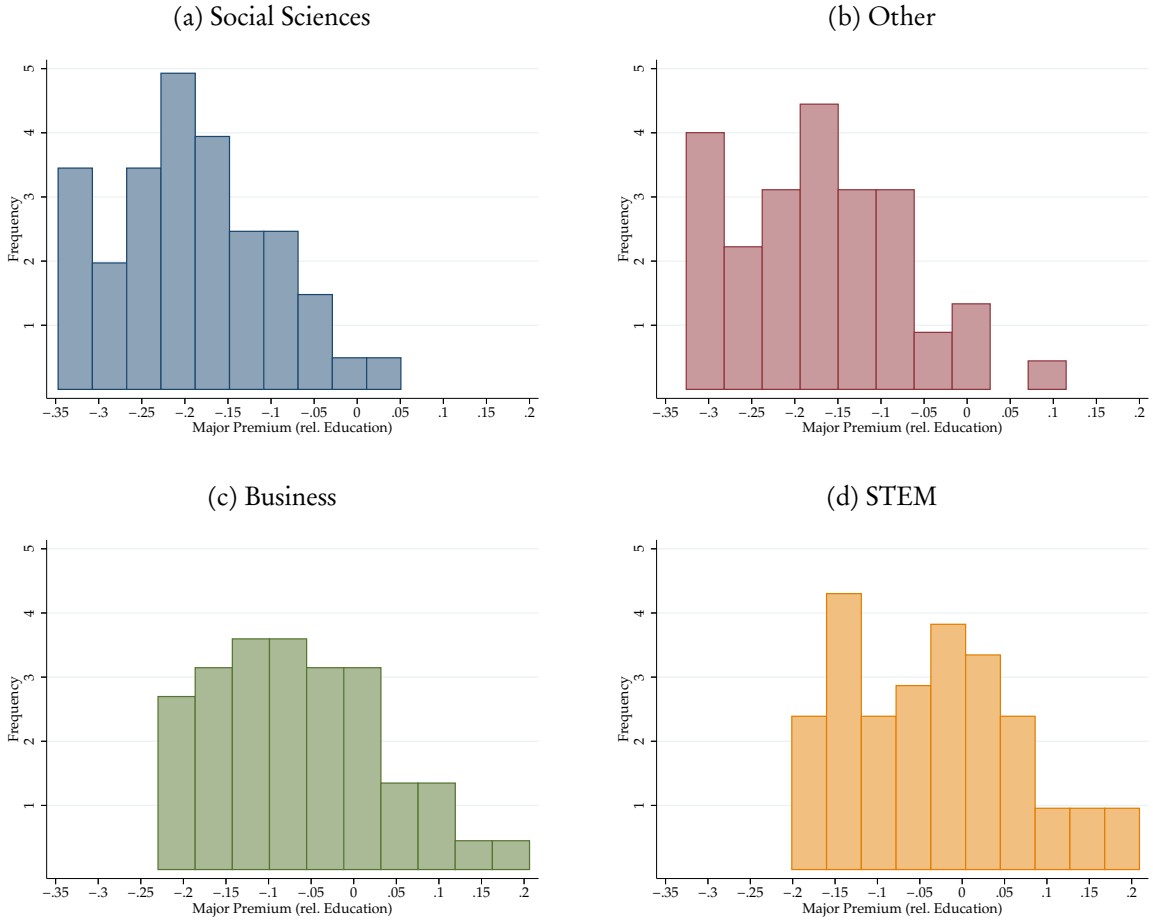
Source: Author's calculations from American Community Survey, 2010-2014.

Figure 2: Major-specific migration distributions across locations



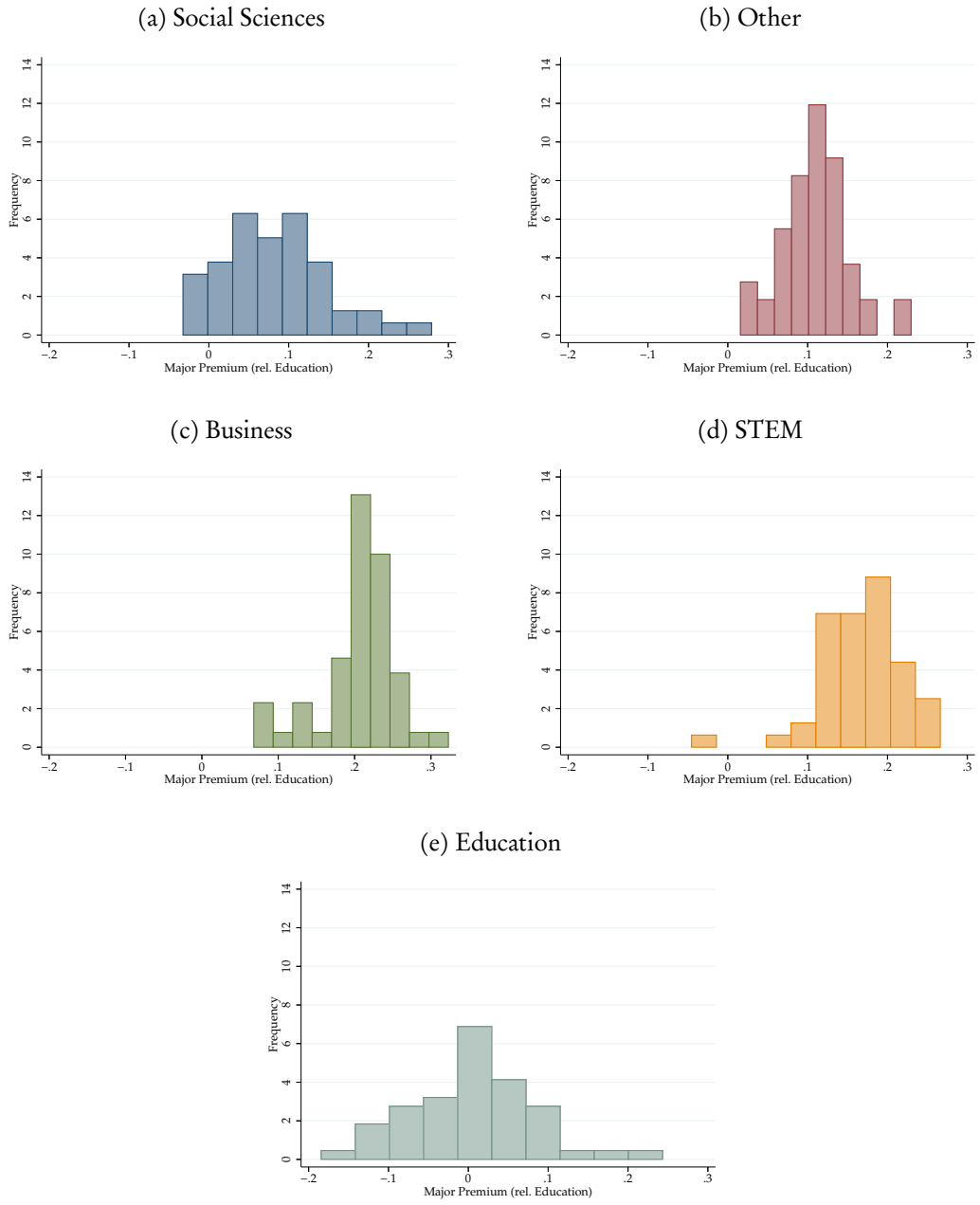
Notes: Above are histograms of the coefficient on major dummies in a linear probability model where "moved away from birth state" is the dependent variable, conditional on residing in a specific U.S. State. Additional controls in the regression include a cubic in potential experience, gender and race dummies, CBSA dummies, and a married dummy. Census population weights are used in the calculations.
Source: Author's calculations from American Community Survey, 2010-2014.

Figure 3: Major-specific occupation relatedness distributions across locations



Notes: Above are histograms of the coefficient on major dummies in a linear probability model where "works in an occupation related to the major" is the dependent variable, conditional on residing in a specific U.S. State. Additional controls in the regression include a cubic in potential experience, gender and race dummies, CBSA dummies, and a married dummy. Census population weights are used in the calculations.
Source: Author's calculations from American Community Survey, 2010-2014.

Figure 4: Distributions of within-location earnings premium for working in an occupation related to one's major



Notes: Above are histograms of the within-location difference in the coefficient on major dummies interacted with an occupation relatedness dummy in a log earnings regression, conditional on residing in a specific U.S. State. Additional controls in the regression include a cubic in potential experience, gender and race dummies, CBSA dummies, and a married dummy. Census population weights are used in the calculations.

Source: Author's calculations from American Community Survey, 2010-2014.

Table 3: Cross-major correlations in outcomes

(a) Log earnings

Major	Soc Sci	Other	Business	STEM
Soc Sci	1.000			
Other	0.822	1.000		
Business	0.794	0.876	1.000	
STEM	0.759	0.797	0.713	1.000

(b) Migration

Major	Soc Sci	Other	Business	STEM
Soc Sci	1.000			
Other	0.690	1.000		
Business	0.648	0.631	1.000	
STEM	0.613	0.739	0.690	1.000

(c) Occupation relatedness propensity

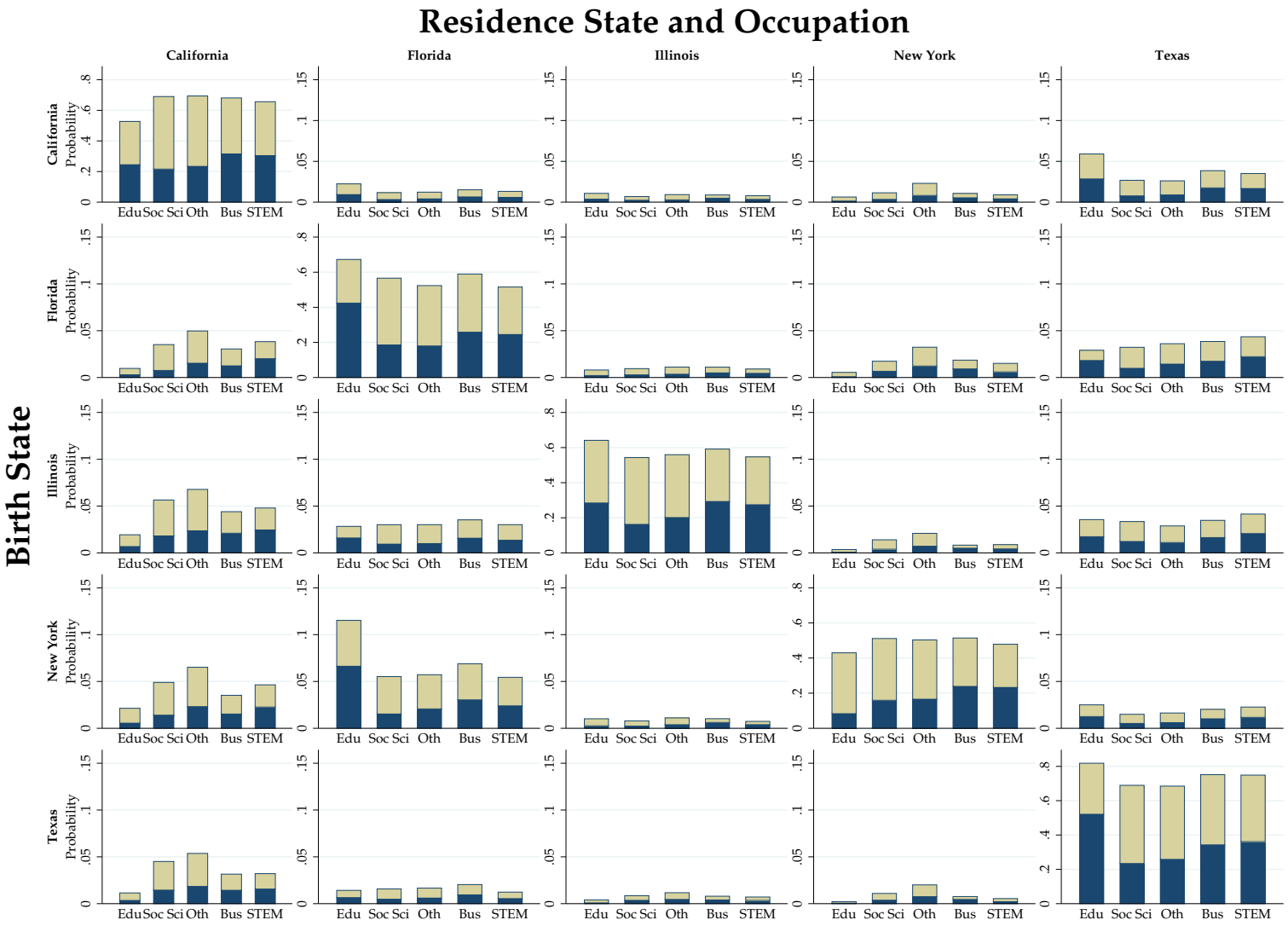
Major	Soc Sci	Other	Business	STEM
Soc Sci	1.000			
Other	0.936	1.000		
Business	0.923	0.938	1.000	
STEM	0.924	0.914	0.947	1.000

(d) Occupation relatedness premium

Major	Edu	Soc Sci	Other	Business	STEM
Education	1.000				
Soc Sci	0.032	1.000			
Other	-0.011	0.102	1.000		
Business	0.031	-0.172	0.190	1.000	
STEM	0.015	-0.200	-0.006	0.143	1.000

Note: This table computes the correlation across majors for various outcomes. Correlations that are close to 1 imply that location characteristics explain the outcome, regardless of major.
Source: Author's calculations from American Community Survey, 2010-2014.

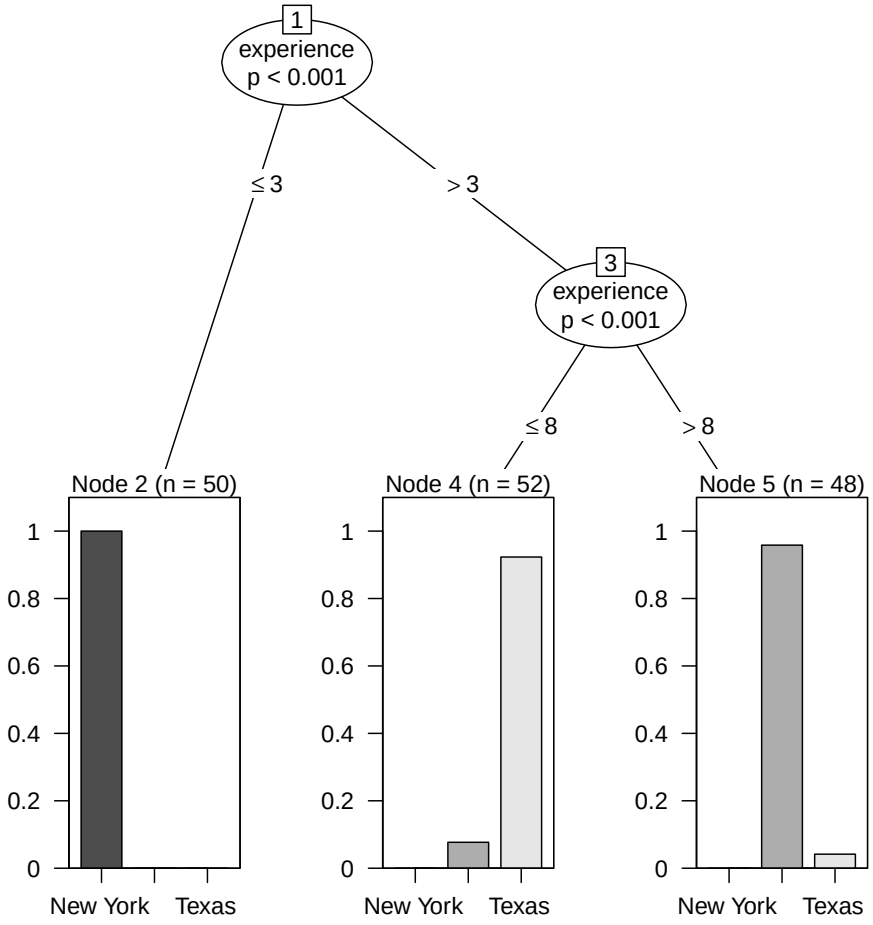
Figure 5: Migration transition matrix by major for the five largest states



Notes: Markov transition matrix probabilities, by major, for five large US states. Light-colored bar segments represent proportion working in an unrelated occupation. Dark-colored bar segments represent proportion working in a related occupation.

Source: Author's calculations from American Community Survey, 2010-2014.

Figure 6: Simple example of tree structure from conditional inference recursive partitioning algorithm



Note: Sample tree output from fictitious data using the algorithm described in Section 5.1.2

Table 4: Summary of cell probabilities of observed decisions: Stayers

(a) Related occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Social Sciences Major	299	23,330	0.4979	0.1907	0.2712	0.7514
Other Major	319	18,744	0.2899	0.0990	0.1881	0.4113
Business Major	324	35,233	0.2966	0.0959	0.1941	0.3904
Education Major	380	59,199	0.3637	0.1016	0.2481	0.4780
STEM Major	367	59,018	0.3640	0.1116	0.2481	0.4821

(b) Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Social Sciences Major	290	9,683	0.2403	0.1071	0.0983	0.3639
Other Major	318	21,338	0.3034	0.1064	0.1999	0.4403
Business Major	334	35,566	0.2865	0.0966	0.1970	0.4051
Education Major	335	40,164	0.2473	0.0862	0.1721	0.3433
STEM Major	327	35,227	0.2331	0.0913	0.1576	0.3433

Note: Estimated decision probabilities and cell structure from the conditional inference recursive partitioning algorithm described in Section 5.1.2. Probabilities correspond to the probability of making the decision that is observed in the data. Source: Author's calculations from American Community Survey, 2010-2014.

Table 5: Summary of cell probabilities of observed decisions: Movers

(a) Related occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Social Sciences Major	527	11,776	0.0215	0.0254	0.0021	0.0487
Other Major	651	14,068	0.0175	0.0232	0.0020	0.0381
Business Major	692	31,387	0.0215	0.0262	0.0025	0.0538
Education Major	721	43,225	0.0213	0.0252	0.0025	0.0477
STEM Major	717	47,119	0.0197	0.0233	0.0025	0.0418

(b) Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Social Sciences Major	536	6,145	0.0128	0.0186	0.0015	0.0269
Other Major	598	16,852	0.0176	0.0224	0.0020	0.0386
Business Major	617	29,847	0.0180	0.0224	0.0020	0.0417
Education Major	649	27,480	0.0144	0.0197	0.0016	0.0298
STEM Major	642	28,447	0.0131	0.0188	0.0015	0.0270

Note: Estimated decision probabilities and cell structure from the conditional inference recursive partitioning algorithm described in Section 5.1.2. Probabilities correspond to the probability of making the decision that is observed in the data. Source: Author's calculations from American Community Survey, 2010-2014.

Table 6: Uncorrected vs. corrected earnings equation estimates for select states

	California		Florida		Illinois		New York		Texas	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
Unrelated occupation										
Education major	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Social sciences major	0.043*	0.120**	0.082***	0.127***	0.030	0.031	0.142***	0.140***	0.044**	0.076**
	(0.023)	(0.051)	(0.024)	(0.028)	(0.024)	(0.035)	(0.026)	(0.031)	(0.021)	(0.035)
Other major	0.017	0.092*	0.048**	0.092***	0.004	0.001	0.091***	0.083***	0.011	0.036
	(0.022)	(0.047)	(0.023)	(0.025)	(0.022)	(0.033)	(0.024)	(0.029)	(0.019)	(0.036)
Business major	0.142***	0.169***	0.147***	0.185***	0.165***	0.141***	0.222***	0.206***	0.114***	0.128***
	(0.023)	(0.044)	(0.023)	(0.026)	(0.022)	(0.020)	(0.025)	(0.044)	(0.019)	(0.033)
STEM major	0.136***	0.165***	0.137***	0.170***	0.177***	0.158***	0.209***	0.185***	0.186***	0.198***
	(0.023)	(0.044)	(0.023)	(0.030)	(0.023)	(0.037)	(0.025)	(0.034)	(0.019)	(0.034)
Related occupation										
Education major	-0.040	0.010	-0.012	0.065	-0.036	0.067	-0.075**	-0.018	0.008	0.009
	(0.027)	(0.094)	(0.024)	(0.118)	(0.025)	(0.100)	(0.036)	(0.113)	(0.019)	(0.078)
Social sciences major	0.121***	0.226**	0.137***	0.270**	0.129***	0.237**	0.208***	0.264**	0.073***	0.181**
	(0.023)	(0.089)	(0.025)	(0.110)	(0.025)	(0.104)	(0.026)	(0.107)	(0.021)	(0.069)
Other major	0.131***	0.238***	0.140***	0.271**	0.131***	0.245**	0.202***	0.250**	0.085***	0.189***
	(0.022)	(0.088)	(0.023)	(0.111)	(0.022)	(0.100)	(0.024)	(0.107)	(0.019)	(0.070)
Business major	0.353***	0.426***	0.361***	0.485***	0.375***	0.481***	0.478***	0.514***	0.345***	0.432***
	(0.022)	(0.088)	(0.022)	(0.106)	(0.021)	(0.103)	(0.024)	(0.107)	(0.018)	(0.074)
STEM major	0.362***	0.432***	0.332***	0.452***	0.299***	0.407***	0.404***	0.434***	0.297***	0.382***
	(0.022)	(0.087)	(0.022)	(0.111)	(0.022)	(0.099)	(0.025)	(0.106)	(0.018)	(0.073)
Married	0.131***	0.108***	0.131***	0.123***	0.131***	0.121***	0.125***	0.097***	0.102***	0.096***
	(0.005)	(0.016)	(0.007)	(0.013)	(0.007)	(0.013)	(0.008)	(0.019)	(0.006)	(0.011)
Female	-0.192***	-0.192***	-0.239***	-0.239***	-0.222***	-0.224***	-0.192***	-0.193***	-0.251***	-0.252***
	(0.005)	(0.014)	(0.007)	(0.015)	(0.007)	(0.021)	(0.007)	(0.021)	(0.005)	(0.017)
Black	-0.160***	-0.153***	-0.158***	-0.150***	-0.182***	-0.159***	-0.280***	-0.256***	-0.205***	-0.209***
	(0.012)	(0.019)	(0.012)	(0.018)	(0.014)	(0.017)	(0.013)	(0.024)	(0.009)	(0.017)
Hispanic	-0.142***	-0.066***	-0.081***	-0.081***	-0.138***	-0.111***	-0.182***	-0.149***	-0.118***	-0.126***
	(0.007)	(0.019)	(0.012)	(0.021)	(0.015)	(0.023)	(0.014)	(0.022)	(0.008)	(0.017)
Other race	-0.112***	-0.039**	-0.089***	-0.108***	0.009	0.024	-0.077***	-0.074***	-0.094***	-0.104***
	(0.007)	(0.017)	(0.022)	(0.033)	(0.019)	(0.027)	(0.015)	(0.023)	(0.015)	(0.021)
Cubic in experience	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CBSA fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wald test for λ terms		5.05		5.96		6.30		14.99		6.31
		[0.000]		[0.000]		[0.000]		[0.000]		[0.000]
R^2	0.244	0.252	0.229	0.234	0.267	0.268	0.231	0.244	0.253	0.259
Observations	58,377	58,377	28,288	28,288	28,697	28,697	34,511	34,511	46,932	46,932

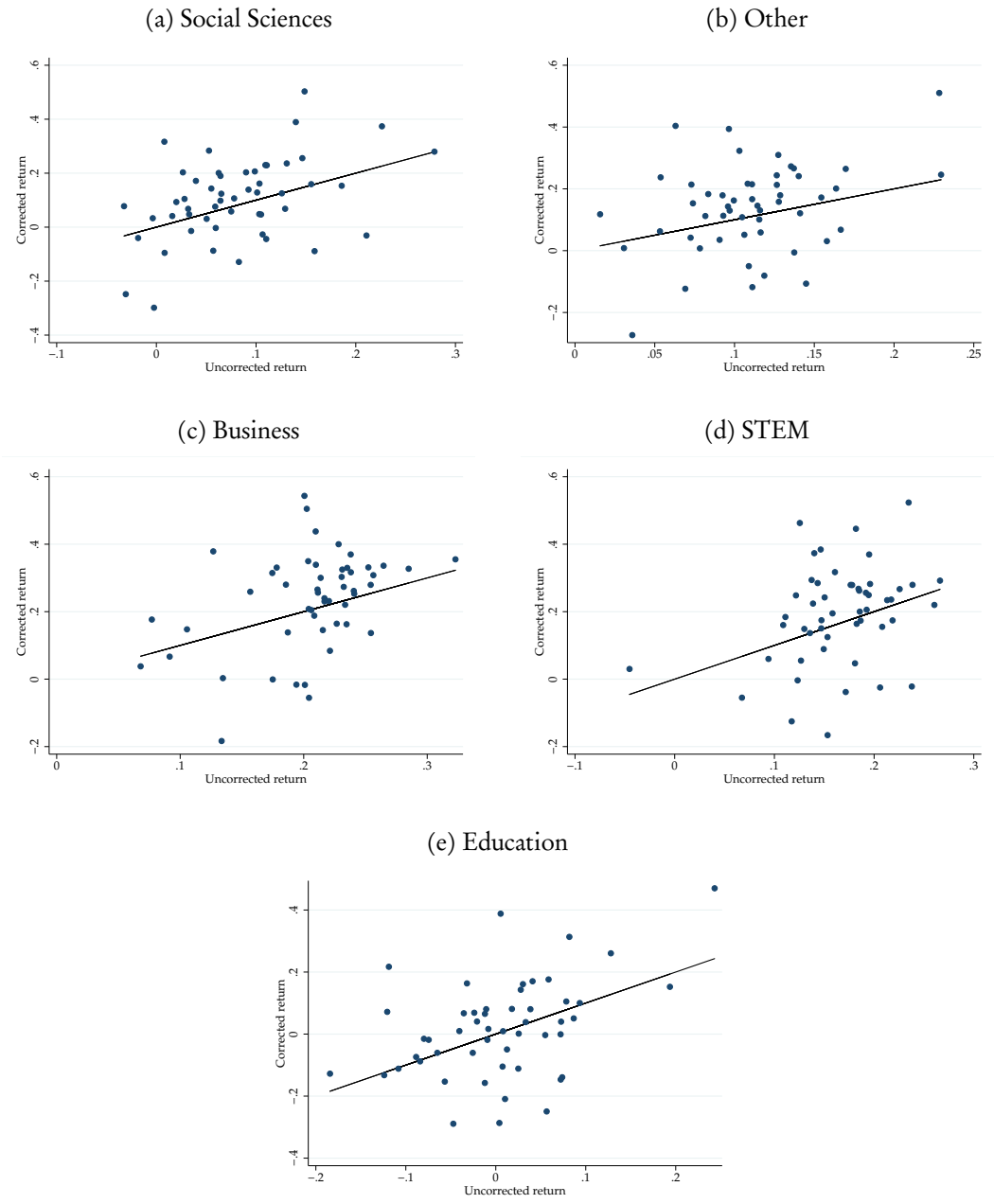
Note: Standard errors are listed below coefficients in parentheses. P-values of statistical tests are listed below test statistics in brackets. *** $p<0.01$; ** $p<0.05$; * $p<0.10$.

Table 7: Aggregate returns to working in related occupation

	Uncorrected			Corrected		
Major	Mean	Std. Dev.	Skewness	Mean	Std. Dev.	Skewness
Education	0.0052	0.0789	0.3192	0.0183	0.1578	0.4484
Social Sciences	0.0814	0.0650	0.6528	0.1038	0.1513	-0.0952
Other	0.1105	0.0420	0.4411	0.1363	0.1434	-0.2149
Business	0.2052	0.0500	-0.8746	0.2322	0.1442	-0.5893
STEM	0.1657	0.0523	-1.0919	0.1915	0.1444	-0.2947

Note: Summary statistics of the 51-location distribution of the return to working in a related occupation for each of the five majors, both with and without selection correction.

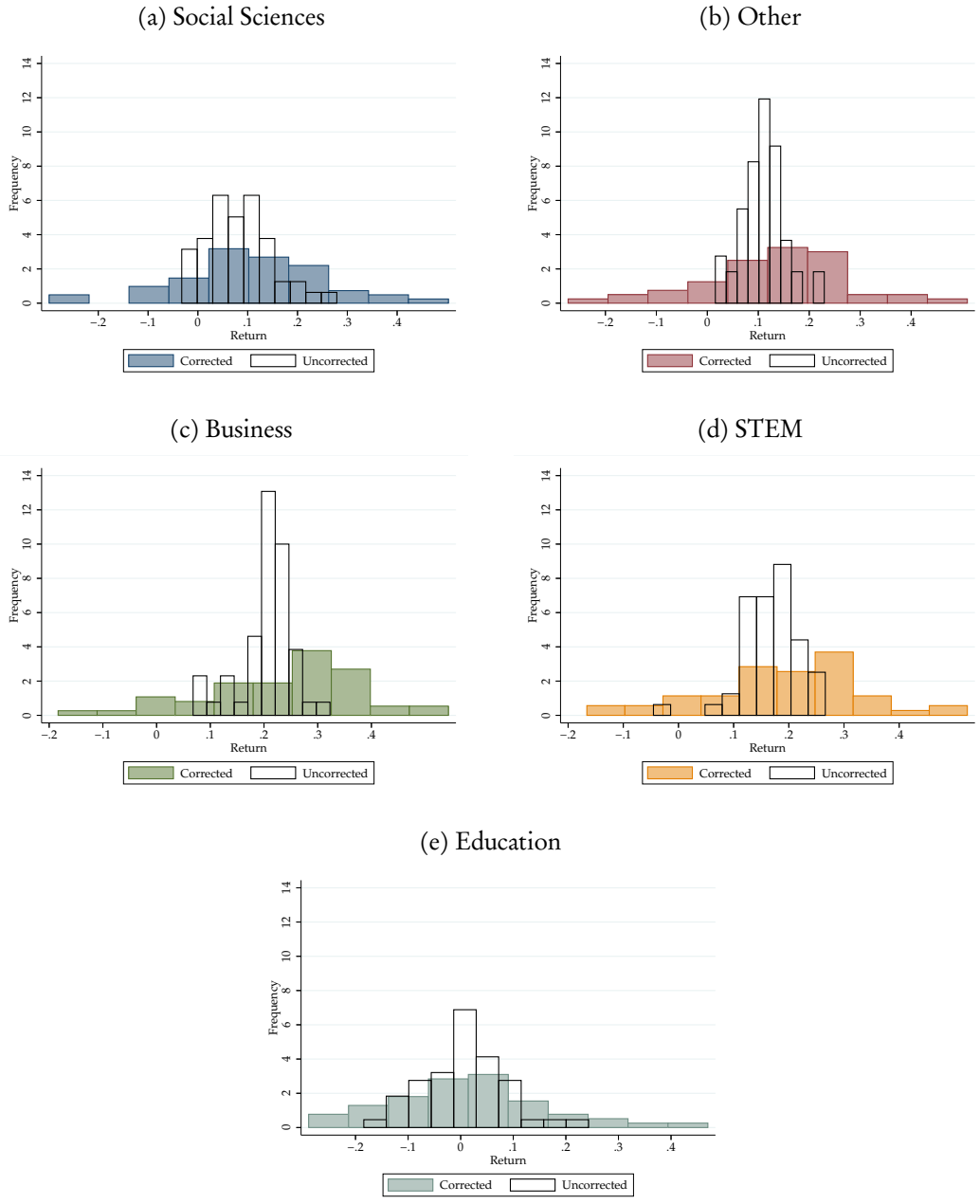
Figure 7: Scatter plots of uncorrected and corrected returns to working in a related occupation



Notes: Scatter plots of return to working in a related occupation, by major. Solid black lines are 45-degree lines. Blue dots are state-specific pairs marking the uncorrected and corrected returns.

Source: Author's calculations from American Community Survey, 2010-2014.

Figure 8: Distributions of corrected returns to working in an occupation related to major



Notes: Above are histograms of the corrected and uncorrected returns to working in a related occupation, by major.

Source: Author's calculations from American Community Survey, 2010-2014.

Table 8: Percent change in returns when correcting for selection

Major	Unrelated occupation			Related occupation			Difference		
	p10	Median	p90	p10	Median	p90	p10	Median	p90
Education	0	0	0	-610.4	-30.7	352.4	-610.4	-30.7	352.4
Social Sciences	-104.6	-3.3	106.2	-207.2	41.8	190.9	-156	52.5	364.8
Other	-81.6	3.4	185.6	-166.9	22.4	154.8	-146	27.1	193.1
Business	-35.9	-.8	33.9	-60.8	7.5	45	-97.8	14.8	85.7
STEM	-19.8	-1.5	25.4	-63	6.4	49.7	-112.1	14.5	114

Note: Summary statistics of the 51-location distribution of the percent change between uncorrected and corrected returns to majors.

A Data Appendix

Table A1: Sample selection details

Criterion	No. obs deleted	Remaining obs.
Respondents in 2010-2014 ACS	—	15,552,144
Drop those without exactly a bachelor's degree	13,498,591	2,053,553
Drop those outside of 22-54 age range	712,627	1,340,926
Drop those with imputed critical variables	279,720	1,061,206
Drop those currently enrolled in school	115,582	945,624
Drop those currently residing in group quarters	6,004	939,620
Drop those not born in the US	136,876	802,744
Drop those with positive annual earnings below $20,000	98,841	703,903
Drop those with annual earnings above $600,000	164	703,739
Drop those with zero annual earnings	109,891	593,848
Final analysis sample	—	593,848

Table A2: Aggregation of the 51 detailed Department of Education majors

Education	*STEM*	*Other*
Primary Education	Agriculture and Agr. Science	Architecture
Secondary Education	All Other Engineering	Area, Ethnic, and Civ. Studies
	Biological Sciences	Art History and Fine Arts
Social Sciences	Chemical Engineering	Commercial Art and Design
Family and Consumer Science	Chemistry	Communications
International Relations	Civil Engineering	Film and Other Arts
Other Social Science	Computer Programming	Foreign Language
Philosophy and Religion	Computer and Info Tech	History
Political Science	Earth and Other Physical Sci	Journalism
Psychology	Electrical Engineering	Leisure Studies
Social Work and HR	Engineering Tech	Letters: Lit, Writing, Other
	Environmental Studies	Music and Speech/Drama
Business	Fitness and Nutrition	Prec. Prod. and Ind. Arts
Accounting	General Science	Protective Services
Business Mgt. and Admin.	Mathematics	Public Admin and Law
Economics	Mechanical Engineering	Public Health
Finance	Medical Tech	
Marketing	Nursing	
Misc. Bus. and Med. Support	Other Med/Health Services	
	Physics	

Note: Aggregation of the 51 detailed Department of Education majors analyzed in Altonji et al. (2016b).

Table A3: List of related majors for each occupation

Occupation	Edu.	Soc. Sci.	Other	Bus.	STEM
Chief executives and public administrators		✓		✓	✓
Financial managers			✓	✓	✓
Human resources and labor relations managers		✓			
Managers and specialists in marketing, advertising, and public relations		✓	✓	✓	✓
Managers in education and related fields		✓			
Managers of medicine and health occupations					✓
Managers of food-serving and lodging establishments			✓		
Managers of service organizations, n.e.c.		✓	✓		
Managers and administrators, n.e.c.	✓	✓	✓	✓	✓
Accountants and auditors			✓	✓	✓
Other financial specialists		✓	✓	✓	✓
Management analysts		✓	✓	✓	✓
Personnel, HR, training, and labor relations specialists		✓	✓	✓	
Inspectors and compliance officers, outside construction			✓		✓
Architects			✓		
Aerospace engineer					✓
Chemical engineers					✓
Civil engineers			✓		✓
Electrical engineer					✓
Industrial engineers					✓
Mechanical engineers					✓
Not-elsewhere-classified engineers					✓
Computer systems analysts and computer scientists		✓	✓	✓	✓
Actuaries					✓
Chemists					✓
Atmospheric and space scientists					✓
Geologists					✓
Physical scientists, n.e.c.					✓
Biological scientists					✓
Foresters and conservation scientists					✓
Registered nurses		✓	✓		✓
Pharmacists					✓
Dietitians and nutritionists					✓
Respiratory therapists					✓
Occupational therapists					✓
Physical therapists					✓
Therapists, n.e.c.					✓
Kindergarten and earlier school teachers	✓	✓			
Primary school teachers	✓	✓	✓		✓
Secondary school teachers	✓		✓		✓
Special education teachers	✓				
Teachers , n.e.c.	✓	✓	✓		
Vocational and educational counselors		✓			
Economists, market researchers, and survey researchers			✓	✓	
Social workers		✓	✓		
Recreation workers					✓
Clergy and religious workers		✓			
Writers and authors			✓		
Designers			✓		
Musician or composer			✓		
Actors, directors, producers			✓		
Art makers: painters, sculptors, craft-artists, and print-makers			✓		
Photographers			✓		
Editors and reporters			✓		
Athletes, sports instructors, and officials					✓
Clinical laboratory technologies and technicians					✓
Dental hygenists				✓	
Radiologic tech specialists					✓
Health technologists and technicians, n.e.c.			✓		✓
Engineering technicians, n.e.c.					✓
Drafters			✓		
Chemical technicians					✓
Airplane pilots and navigators			✓		
Air traffic controllers			✓		
Computer software developers				✓	✓
Legal assistants, paralegals, legal support, etc		✓	✓		
Supervisors and proprietors of sales jobs	✓	✓	✓	✓	✓
Insurance sales occupations				✓	
Financial services sales occupations				✓	
Salespersons, n.e.c.	✓	✓	✓	✓	✓
Retail sales clerks		✓	✓	✓	✓
Office supervisors		✓	✓	✓	
Secretaries		✓	✓	✓	
Transportation ticket and reservation agents			✓		
Customer service reps, investigators and adjusters, except insurance		✓	✓	✓	
Administrative support jobs, n.e.c.			✓		
Fire fighting, prevention, and inspection			✓		✓
Police, detectives, and private investigators		✓	✓		✓
Other law enforcement: sheriffs, bailiffs, correctional institution officers			✓		
Guards, watchmen, doorkeepers			✓		
Waiter/waitress			✓		
Cooks, variously defined			✓		
Welfare service aides		✓	✓		
Child care workers		✓			
Farmers (owners and tenants)					✓
Farm workers					✓
Supervisors of agricultural occupations					✓
Gardeners and groundskeepers					✓
Supervisors of construction work			✓		
Production supervisors or foremen					✓
Military		✓	✓		✓

Note: Occupations not related to any college major are excluded from this table.

Table A4: Predictive performance of various algorithms

Performance Criterion	Classification algorithm Logit	Bin	Tree
Training set performance:			
Accuracy	35.34%	35.16%	34.72%
Kappa	33.95%	33.76%	33.25%
Test set performance:			
Accuracy	35.22%	34.51%	33.91%
Kappa	33.82%	33.09%	32.42%

Note: "Logit" refers to a multinomial logit; "Bin" refers to a simple bin estimator; "Tree" refers to the conditional inference tree classification algorithm detailed in Section 5.1.2. I estimate each algorithm on a subset of the 2010-2014 ACS sample included in this paper and compute predictive performance out-of-sample using a holdout sample. To measure predictive performance, I compute the predicted alternative, defined as the alternative with the largest predicted probability. Predictive performance is measured via a multi-dimensional confusion matrix using two related but separate metrics: Accuracy and Kappa.

$\text{Accuracy} = \frac{\text{number of correctly classified predictions}}{\text{number of predictions}}$.

$\text{Kappa} = \frac{\text{Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}}$.

Expected Accuracy is defined as $\text{Expected Accuracy} = \sum_{j=1}^{J} \left[\left(\sum_i d_{ij} \right) \left(\sum_i p_{ij} \right) \right] / N^J$, where d_{ij} represents the observed class for observation i in the data, p_{ij} represents the predicted class for observation i, and N represents the total number of observations. The Kappa statistic is meant to capture predictive performance net of guessing. For example, the Kappa statistic penalizes strategies that would predict that all observations belong to one class.

Table A5: Return to STEM majors in unrelated occupation, by state (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms	State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
Alabama	0.221	0.218	0.070	19.223	Montana	0.070	0.090	0.576	4430.798
	(0.036)	(0.046)	[0.791]	[0.000]		(0.057)	(0.039)	[0.448]	[0.000]
Alaska	0.094	-0.097	5.854	74.969	Nebraska	0.128	0.121	0.222	148.770
	(0.126)	(0.053)	[0.016]	[0.000]		(0.042)	(0.048)	[0.638]	[0.000]
Arizona	0.154	0.151	0.151	10.662	Nevada	0.143	0.125	2.110	2440.229
	(0.038)	(0.049)	[0.697]	[0.000]		(0.070)	(0.088)	[0.146]	[0.000]
Arkansas	0.084	0.117	0.827	34.807	New Hampshire	0.303	0.281	3.436	2123.841
	(0.048)	(0.049)	[0.363]	[0.000]		(0.059)	(0.081)	[0.064]	[0.000]
California	0.136	0.165	2.599	5.049	New Jersey	0.190	0.179	0.184	7.013
	(0.023)	(0.044)	[0.107]	[0.000]		(0.031)	(0.050)	[0.668]	[0.000]
Colorado	0.160	0.146	0.249	7.161	New Mexico	0.205	0.176	2.274	23.519
	(0.037)	(0.035)	[0.618]	[0.000]		(0.086)	(0.088)	[0.132]	[0.000]
Connecticut	0.133	0.111	0.826	38.673	New York	0.209	0.185	0.649	14.986
	(0.045)	(0.049)	[0.364]	[0.000]		(0.025)	(0.044)	[0.421]	[0.000]
Delaware	0.138	0.153	0.204	4332.845	North Carolina	0.192	0.228	2.524	15.461
	(0.085)	(0.114)	[0.652]	[0.000]		(0.026)	(0.026)	[0.112]	[0.000]
District of Columbia	0.237	0.219	0.302	29.314	North Dakota	0.075	0.160	0.938	61344.370
	(0.115)	(0.064)	[0.582]	[0.000]		(0.064)	(0.094)	[0.333]	[0.000]
Florida	0.137	0.170	5.606	5.956	Ohio	0.232	0.251	0.987	8.114
	(0.023)	(0.026)	[0.018]	[0.000]		(0.022)	(0.014)	[0.321]	[0.000]
Georgia	0.194	0.219	1.587	15.943	Oklahoma	0.138	0.101	1.463	13.301
	(0.029)	(0.032)	[0.208]	[0.000]		(0.038)	(0.080)	[0.226]	[0.000]
Hawaii	0.277	0.280	0.012	47.869	Oregon	0.092	0.100	0.425	24.295
	(0.075)	(0.067)	[0.911]	[0.000]		(0.047)	(0.036)	[0.515]	[0.000]
Idaho	0.218	0.197	0.599	1522.202	Pennsylvania	0.277	0.259	4.566	7.785
	(0.069)	(0.091)	[0.439]	[0.000]		(0.020)	(0.025)	[0.033]	[0.000]
Illinois	0.177	0.158	1.947	6.298	Rhode Island	0.170	0.136	4.018	35973.400
	(0.022)	(0.020)	[0.163]	[0.000]		(0.082)	(0.060)	[0.045]	[0.000]
Indiana	0.135	0.181	8.641	86.113	South Carolina	0.204	0.208	0.132	44.641
	(0.029)	(0.029)	[0.003]	[0.000]		(0.036)	(0.043)	[0.716]	[0.000]
Iowa	0.198	0.175	2.060	1099.694	South Dakota	0.165	0.206	0.423	379.184
	(0.034)	(0.044)	[0.151]	[0.000]		(0.069)	(0.108)	[0.515]	[0.000]
Kansas	0.170	0.202	0.353	2188.396	Tennessee	0.229	0.236	1.207	29.565
	(0.039)	(0.050)	[0.552]	[0.000]		(0.032)	(0.036)	[0.272]	[0.000]
Kentucky	0.222	0.132	10.895	32.843	Texas	0.186	0.198	0.154	5.095
	(0.040)	(0.046)	[0.001]	[0.000]		(0.019)	(0.033)	[0.695]	[0.000]
Louisiana	0.221	0.194	1.554	44.647	Utah	0.200	0.261	1.790	33.019
	(0.043)	(0.039)	[0.213]	[0.000]		(0.051)	(0.055)	[0.181]	[0.000]
Maine	0.194	0.174	32.191	8871.024	Vermont	0.213	0.232	2.128	168.104
	(0.070)	(0.010)	[0.000]	[0.000]		(0.077)	(0.154)	[0.145]	[0.000]
Maryland	0.153	0.169	1.796	8.368	Virginia	0.221	0.243	0.771	9.822
	(0.037)	(0.045)	[0.180]	[0.000]		(0.029)	(0.040)	[0.380]	[0.000]
Massachusetts	0.239	0.218	0.660	12.339	Washington	0.187	0.185	0.034	7.577
	(0.032)	(0.034)	[0.416]	[0.000]		(0.037)	(0.036)	[0.854]	[0.000]
Michigan	0.182	0.149	3.388	16.788	West Virginia	0.122	0.063	0.509	2715.274
	(0.026)	(0.014)	[0.066]	[0.000]		(0.061)	(0.081)	[0.475]	[0.000]
Minnesota	0.287	0.309	0.824	16.720	Wisconsin	0.193	0.189	0.147	41.088
	(0.029)	(0.035)	[0.364]	[0.000]		(0.029)	(0.035)	[0.702]	[0.000]
Mississippi	0.167	0.130	0.655	587.065	Wyoming	0.283	0.303	0.402	14381.270
	(0.045)	(0.051)	[0.418]	[0.000]		(0.088)	(0.038)	[0.526]	[0.000]
Missouri	0.191	0.203	1.895	20.119					
	(0.031)	(0.033)	[0.169]	[0.000]					

Table A6: Return to STEM majors in related occupation, by state (uncorrected and corrected)

State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms	State	Uncorrected STEM Return	Corrected STEM Return	χ^2 Test for Difference	Wald Test for Correction Terms
Alabama	0.429	0.373	0.229	19.223	Montana	0.256	0.290	1.363	4430.798
	(0.036)	(0.127)	[0.632]	[0.000]		(0.055)	(0.039)	[0.243]	[0.000]
Alaska	0.248	0.028	6.147	74.969	Nebraska	0.281	-0.045	9.854	148.770
	(0.124)	(0.096)	[0.013]	[0.000]		(0.040)	(0.122)	[0.002]	[0.000]
Arizona	0.367	0.385	0.293	10.662	Nevada	0.382	0.404	0.015	2440.229
	(0.037)	(0.048)	[0.588]	[0.000]		(0.069)	(0.194)	[0.904]	[0.000]
Arkansas	0.234	0.206	0.165	34.807	New Hampshire	0.450	0.456	0.001	2123.841
	(0.046)	(0.071)	[0.685]	[0.000]		(0.057)	(0.212)	[0.971]	[0.000]
California	0.362	0.432	0.839	5.049	New Jersey	0.326	0.316	0.007	7.013
	(0.022)	(0.088)	[0.360]	[0.000]		(0.030)	(0.121)	[0.934]	[0.000]
Colorado	0.347	0.320	0.079	7.161	New Mexico	0.471	0.468	0.002	23.519
	(0.036)	(0.111)	[0.778]	[0.000]		(0.083)	(0.111)	[0.967]	[0.000]
Connecticut	0.311	0.390	0.489	38.673	New York	0.404	0.434	0.067	14.986
	(0.044)	(0.130)	[0.485]	[0.000]		(0.024)	(0.107)	[0.796]	[0.000]
Delaware	0.263	0.616	3.939	4332.845	North Carolina	0.319	0.283	0.224	15.461
	(0.081)	(0.179)	[0.047]	[0.000]		(0.025)	(0.091)	[0.636]	[0.000]
District of Columbia	0.191	0.249	0.342	29.314	North Dakota	0.257	0.606	15.174	61344.370
	(0.115)	(0.134)	[0.559]	[0.000]		(0.058)	(0.103)	[0.000]	[0.000]
Florida	0.332	0.452	1.096	5.956	Ohio	0.383	0.493	0.591	8.114
	(0.022)	(0.106)	[0.295]	[0.000]		(0.022)	(0.155)	[0.442]	[0.000]
Georgia	0.324	0.368	0.327	15.943	Oklahoma	0.333	0.471	1.785	13.301
	(0.028)	(0.083)	[0.568]	[0.000]		(0.037)	(0.145)	[0.181]	[0.000]
Hawaii	0.462	0.542	0.672	47.869	Oregon	0.330	0.078	1.995	24.295
	(0.075)	(0.140)	[0.412]	[0.000]		(0.045)	(0.212)	[0.158]	[0.000]
Idaho	0.479	0.417	0.875	1522.202	Pennsylvania	0.420	0.544	1.179	7.785
	(0.067)	(0.092)	[0.350]	[0.000]		(0.019)	(0.117)	[0.278]	[0.000]
Illinois	0.299	0.407	1.298	6.298	Rhode Island	0.279	0.297	0.178	35973.400
	(0.021)	(0.103)	[0.254]	[0.000]		(0.079)	(0.063)	[0.673]	[0.000]
Indiana	0.327	0.437	0.572	86.113	South Carolina	0.384	0.255	1.614	44.641
	(0.028)	(0.145)	[0.450]	[0.000]		(0.035)	(0.128)	[0.204]	[0.000]
Iowa	0.370	0.137	1.365	1099.694	South Dakota	0.259	0.266	0.007	379.184
	(0.033)	(0.195)	[0.243]	[0.000]		(0.067)	(0.081)	[0.933]	[0.000]
Kansas	0.347	0.482	2.093	2188.396	Tennessee	0.352	0.233	2.002	29.565
	(0.038)	(0.102)	[0.148]	[0.000]		(0.031)	(0.097)	[0.157]	[0.000]
Kentucky	0.361	0.356	0.001	32.843	Texas	0.297	0.382	1.486	5.095
	(0.039)	(0.213)	[0.982]	[0.000]		(0.018)	(0.074)	[0.223]	[0.000]
Louisiana	0.289	0.139	1.480	44.647	Utah	0.358	0.456	0.652	33.019
	(0.042)	(0.132)	[0.224]	[0.000]		(0.050)	(0.111)	[0.420]	[0.000]
Maine	0.340	0.558	3.225	8871.024	Vermont	0.330	0.107	1.127	168.104
	(0.067)	(0.125)	[0.073]	[0.000]		(0.074)	(0.220)	[0.288]	[0.000]
Maryland	0.336	0.333	0.001	8.368	Virginia	0.359	0.537	6.016	9.822
	(0.036)	(0.099)	[0.971]	[0.000]		(0.028)	(0.091)	[0.014]	[0.000]
Massachusetts	0.386	0.369	0.030	12.339	Washington	0.405	0.421	0.063	7.577
	(0.031)	(0.109)	[0.864]	[0.000]		(0.036)	(0.073)	[0.803]	[0.000]
Michigan	0.342	0.466	0.732	16.788	West Virginia	0.357	0.587	0.975	2715.274
	(0.025)	(0.160)	[0.392]	[0.000]		(0.059)	(0.250)	[0.323]	[0.000]
Minnesota	0.472	0.576	1.349	16.720	Wisconsin	0.412	0.363	0.217	41.088
	(0.027)	(0.100)	[0.245]	[0.000]		(0.028)	(0.108)	[0.642]	[0.000]
Mississippi	0.373	0.105	3.595	587.065	Wyoming	0.423	0.677	7.204	14381.270
	(0.044)	(0.151)	[0.058]	[0.000]		(0.088)	(0.086)	[0.007]	[0.000]
Missouri	0.384	0.409	0.047	20.119					
	(0.031)	(0.118)	[0.828]	[0.000]					

Table A7: Return to Business majors in unrelated occupation, by state (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms	State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
Alabama	0.167	0.185	3.261	19.223	Montana	0.083	0.078	0.017	4430.798
	(0.039)	(0.056)	[0.071]	[0.000]		(0.060)	(0.039)	[0.898]	[0.000]
Alaska	0.134	-0.050	8.310	74.969	Nebraska	0.157	0.151	0.207	148.770
	(0.123)	(0.070)	[0.004]	[0.000]		(0.043)	(0.059)	[0.649]	[0.000]
Arizona	0.144	0.140	0.057	10.662	Nevada	0.167	0.145	2.234	2440.229
	(0.038)	(0.051)	[0.811]	[0.000]		(0.069)	(0.093)	[0.135]	[0.000]
Arkansas	0.058	0.090	0.682	34.807	New Hampshire	0.257	0.261	0.607	2123.841
	(0.051)	(0.041)	[0.409]	[0.000]		(0.059)	(0.066)	[0.436]	[0.000]
California	0.142	0.169	3.192	5.049	New Jersey	0.190	0.169	0.230	7.013
	(0.022)	(0.047)	[0.074]	[0.000]		(0.031)	(0.054)	[0.631]	[0.000]
Colorado	0.126	0.104	0.926	7.161	New Mexico	0.190	0.170	0.624	23.519
	(0.037)	(0.038)	[0.336]	[0.000]		(0.087)	(0.078)	[0.430]	[0.000]
Connecticut	0.126	0.115	0.177	38.673	New York	0.222	0.206	0.256	14.986
	(0.045)	(0.058)	[0.674]	[0.000]		(0.024)	(0.029)	[0.613]	[0.000]
Delaware	0.062	0.072	0.084	4332.845	North Carolina	0.120	0.153	4.507	15.461
	(0.084)	(0.075)	[0.772]	[0.000]		(0.026)	(0.037)	[0.034]	[0.000]
District of Columbia	0.152	0.126	0.467	29.314	North Dakota	0.138	0.234	1.552	61344.370
	(0.113)	(0.085)	[0.494]	[0.000]		(0.065)	(0.089)	[0.213]	[0.000]
Florida	0.147	0.185	7.753	5.956	Ohio	0.217	0.232	0.932	8.114
	(0.023)	(0.025)	[0.005]	[0.000]		(0.023)	(0.027)	[0.334]	[0.000]
Georgia	0.108	0.137	2.025	15.943	Oklahoma	0.085	0.055	1.004	13.301
	(0.030)	(0.042)	[0.155]	[0.000]		(0.040)	(0.077)	[0.316]	[0.000]
Hawaii	0.151	0.166	0.315	47.869	Oregon	0.067	0.075	0.352	24.295
	(0.076)	(0.059)	[0.574]	[0.000]		(0.046)	(0.056)	[0.553]	[0.000]
Idaho	0.279	0.232	1.571	1522.202	Pennsylvania	0.208	0.195	2.087	7.785
	(0.070)	(0.094)	[0.210]	[0.000]		(0.020)	(0.027)	[0.149]	[0.000]
Illinois	0.165	0.141	3.588	6.298	Rhode Island	0.056	0.050	0.089	35973.400
	(0.022)	(0.033)	[0.058]	[0.000]		(0.080)	(0.060)	[0.765]	[0.000]
Indiana	0.111	0.161	3.177	86.113	South Carolina	0.112	0.124	0.983	44.641
	(0.029)	(0.031)	[0.075]	[0.000]		(0.037)	(0.071)	[0.321]	[0.000]
Iowa	0.181	0.163	1.133	1099.694	South Dakota	0.137	0.191	0.351	379.184
	(0.036)	(0.057)	[0.287]	[0.000]		(0.073)	(0.065)	[0.554]	[0.000]
Kansas	0.185	0.223	0.413	2188.396	Tennessee	0.149	0.161	0.572	29.565
	(0.040)	(0.074)	[0.520]	[0.000]		(0.033)	(0.039)	[0.449]	[0.000]
Kentucky	0.163	0.082	9.861	32.843	Texas	0.114	0.128	0.234	5.095
	(0.041)	(0.039)	[0.002]	[0.000]		(0.019)	(0.036)	[0.628]	[0.000]
Louisiana	0.068	0.050	0.744	44.647	Utah	0.177	0.237	1.605	33.019
	(0.043)	(0.046)	[0.388]	[0.000]		(0.051)	(0.050)	[0.205]	[0.000]
Maine	0.148	0.166	4.562	8871.024	Vermont	0.073	0.010	3.908	168.104
	(0.068)	(0.011)	[0.033]	[0.000]		(0.074)	(0.112)	[0.048]	[0.000]
Maryland	0.131	0.155	3.740	8.368	Virginia	0.200	0.232	1.335	9.822
	(0.037)	(0.044)	[0.053]	[0.000]		(0.029)	(0.046)	[0.248]	[0.000]
Massachusetts	0.201	0.190	0.241	12.339	Washington	0.158	0.156	0.010	7.577
	(0.031)	(0.032)	[0.623]	[0.000]		(0.036)	(0.037)	[0.919]	[0.000]
Michigan	0.165	0.146	2.997	16.788	West Virginia	0.148	0.066	1.333	2715.274
	(0.027)	(0.018)	[0.083]	[0.000]		(0.063)	(0.070)	[0.248]	[0.000]
Minnesota	0.252	0.264	0.538	16.720	Wisconsin	0.234	0.230	0.066	41.088
	(0.029)	(0.027)	[0.463]	[0.000]		(0.030)	(0.039)	[0.797]	[0.000]
Mississippi	0.136	0.085	0.954	587.065	Wyoming	0.130	0.202	2.013	14381.270
	(0.049)	(0.071)	[0.329]	[0.000]		(0.086)	(0.071)	[0.156]	[0.000]
Missouri	0.190	0.204	2.725	20.119					
	(0.032)	(0.039)	[0.099]	[0.000]					

Table A8: Return to Business majors in related occupation, by state (uncorrected and corrected)

State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms	State	Uncorrected Business Return	Corrected Business Return	χ^2 Test for Difference	Wald Test for Correction Terms
Alabama	0.382	0.330	0.199	19.223	Montana	0.336	0.410	4.986	4430.798
	(0.039)	(0.122)	[0.655]	[0.000]		(0.063)	(0.036)	[0.026]	[0.000]
Alaska	0.368	0.171	3.535	74.969	Nebraska	0.291	-0.032	9.595	148.770
	(0.127)	(0.098)	[0.060]	[0.000]		(0.043)	(0.120)	[0.002]	[0.000]
Arizona	0.361	0.380	0.367	10.662	Nevada	0.452	0.473	0.012	2440.229
	(0.038)	(0.049)	[0.545]	[0.000]		(0.070)	(0.201)	[0.912]	[0.000]
Arkansas	0.285	0.255	0.202	34.807	New Hampshire	0.477	0.492	0.008	2123.841
	(0.051)	(0.079)	[0.653]	[0.000]		(0.061)	(0.217)	[0.928]	[0.000]
California	0.353	0.426	0.891	5.049	New Jersey	0.431	0.422	0.006	7.013
	(0.022)	(0.088)	[0.345]	[0.000]		(0.030)	(0.120)	[0.937]	[0.000]
Colorado	0.332	0.309	0.062	7.161	New Mexico	0.296	0.318	0.080	23.519
	(0.037)	(0.105)	[0.804]	[0.000]		(0.086)	(0.117)	[0.777]	[0.000]
Connecticut	0.361	0.445	0.565	38.673	New York	0.478	0.514	0.097	14.986
	(0.045)	(0.134)	[0.452]	[0.000]		(0.024)	(0.107)	[0.756]	[0.000]
Delaware	0.263	0.616	3.776	4332.845	North Carolina	0.355	0.316	0.253	15.461
	(0.086)	(0.187)	[0.052]	[0.000]		(0.026)	(0.093)	[0.615]	[0.000]
District of Columbia	0.229	0.303	0.490	29.314	North Dakota	0.265	0.612	15.561	61344.370
	(0.113)	(0.136)	[0.484]	[0.000]		(0.066)	(0.122)	[0.000]	[0.000]
Florida	0.361	0.485	1.107	5.956	Ohio	0.402	0.512	0.582	8.114
	(0.023)	(0.111)	[0.293]	[0.000]		(0.023)	(0.144)	[0.446]	[0.000]
Georgia	0.348	0.398	0.422	15.943	Oklahoma	0.313	0.455	1.771	13.301
	(0.030)	(0.079)	[0.516]	[0.000]		(0.041)	(0.133)	[0.183]	[0.000]
Hawaii	0.415	0.502	0.779	47.869	Oregon	0.271	0.019	1.980	24.295
	(0.078)	(0.115)	[0.378]	[0.000]		(0.046)	(0.214)	[0.159]	[0.000]
Idaho	0.347	0.270	1.242	1522.202	Pennsylvania	0.446	0.565	1.069	7.785
	(0.072)	(0.091)	[0.265]	[0.000]		(0.020)	(0.120)	[0.301]	[0.000]
Illinois	0.375	0.481	1.249	6.298	Rhode Island	0.379	0.406	0.340	35973.400
	(0.022)	(0.100)	[0.264]	[0.000]		(0.081)	(0.060)	[0.560]	[0.000]
Indiana	0.322	0.427	0.527	86.113	South Carolina	0.334	0.208	1.592	44.641
	(0.029)	(0.150)	[0.468]	[0.000]		(0.037)	(0.130)	[0.207]	[0.000]
Iowa	0.375	0.147	1.302	1099.694	South Dakota	0.229	0.257	0.100	379.184
	(0.037)	(0.193)	[0.254]	[0.000]		(0.074)	(0.089)	[0.752]	[0.000]
Kansas	0.341	0.482	2.255	2188.396	Tennessee	0.403	0.297	1.607	29.565
	(0.041)	(0.097)	[0.133]	[0.000]		(0.033)	(0.096)	[0.205]	[0.000]
Kentucky	0.401	0.399	0.000	32.843	Texas	0.345	0.432	1.553	5.095
	(0.041)	(0.210)	[0.992]	[0.000]		(0.019)	(0.070)	[0.213]	[0.000]
Louisiana	0.203	0.053	1.503	44.647	Utah	0.409	0.510	0.681	33.019
	(0.044)	(0.129)	[0.220]	[0.000]		(0.051)	(0.115)	[0.409]	[0.000]
Maine	0.358	0.604	4.042	8871.024	Vermont	0.247	0.010	1.589	168.104
	(0.071)	(0.146)	[0.044]	[0.000]		(0.077)	(0.238)	[0.208]	[0.000]
Maryland	0.340	0.343	0.002	8.368	Virginia	0.404	0.582	5.956	9.822
	(0.037)	(0.098)	[0.965]	[0.000]		(0.029)	(0.092)	[0.015]	[0.000]
Massachusetts	0.405	0.398	0.005	12.339	Washington	0.412	0.436	0.136	7.577
	(0.031)	(0.108)	[0.944]	[0.000]		(0.036)	(0.072)	[0.712]	[0.000]
Michigan	0.340	0.461	0.700	16.788	West Virginia	0.350	0.571	0.914	2715.274
	(0.027)	(0.150)	[0.403]	[0.000]		(0.064)	(0.244)	[0.339]	[0.000]
Minnesota	0.483	0.589	1.353	16.720	Wisconsin	0.421	0.369	0.243	41.088
	(0.028)	(0.098)	[0.245]	[0.000]		(0.030)	(0.105)	[0.622]	[0.000]
Mississippi	0.337	0.068	3.545	587.065	Wyoming	0.308	0.533	5.816	14381.270
	(0.049)	(0.143)	[0.060]	[0.000]		(0.093)	(0.075)	[0.016]	[0.000]
Missouri	0.407	0.434	0.054	20.119					
	(0.032)	(0.108)	[0.816]	[0.000]					