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

Tyler Ransom†

University of Oklahoma

February 7, 2021

Abstract

I examine the extent to which the monetary returns to college majors are influenced by selective migration and occupational choice across locations in the US. To quantify the role of selection, I develop and estimate an extended Roy model of migration, occupational choice, and earnings where, upon completing their education, individuals choose a location in which to live and an occupation in which to work. 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 OLS estimates of the return to business and STEM majors relative to education majors are biased upward by 15% at the median. Selection is strongest in locations in the Northeastern US.

JEL Classification: I2, J3, R1

Keywords: College major, migration, occupation, Roy model

*I would like to thank Peter Arcidiacono, Esteban Aucejo, Pat Bayer, Rob Garlick, Arnaud Maurel, Ronni Pavan, Kevin Stange and various conference and seminar participants 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: Department of Economics, University of Oklahoma, 308 Cate Center Dr. 158 CCD1, Norman, Oklahoma, USA 73072. E-mail: *ransom@ou.edu*. Telephone: +1 405 325 2861.

1 Introduction

One lesser-known characteristic of the US labor market is that the wage returns to different college majors are highly heterogeneous across space. For example, among men in the 2010–2019 American Community Survey, the return to STEM and business majors each range from about 23% to over 47%, relative to education majors.[1] While much work has examined sorting of majors into occupations, occupational sorting does little to narrow this gap: the return to a STEM major in a STEM occupation relative to a STEM major in a non-STEM occupation ranges from 11% in Pennsylvania to 43% in Washington, DC, with a similar range for other majors. This broad range in returns to majors and occupations suggests that post-college migration, and in particular its interaction with post-college occupational choice, might be a significant driver of the observed spatial variation in earnings.

The objective of this paper is to uncover the extent to which selection into residence location and occupation biases the observed monetary returns to college majors (relative to education majors). Aside from Winters (2017), this is the first paper to examine the spatial dimension of college major and occupation decisions.[2]

Understanding the true returns to human capital investments is important because students base these investment decisions in part on expected earnings (Beffy, Fougère, and Maurel, 2012; Wiswall and Zafar, 2015). A student might choose differently if an observed earnings premium in a particular major is overstated due to selectivity of post-college migration or occupation decisions.

Using data on male college graduates from the 2010–2019 American Community Survey, I document substantial differences in earnings, occupational choice, and locational choice across college majors. These differences provide support for the existence of different location-occupation complementarities for different majors.

[1] Returns calculated using a Mincerian regression of log earnings on a cubic in potential experience, demographic indicators, and MSA fixed effects.

[2] Winters (2017) examines the migration response of different college majors to birth-state earnings shocks to workers in the same major.

As an example, I show that STEM and business majors earn the highest returns to and are much more likely to work in occupations related to their major. However, business majors are much less likely to live outside their state of birth. These results are consistent with a model where college graduates have preferences for working in an occupation related to their field of study, but where occupational concentration varies across space.

Additional evidence on the importance of location and occupation for college majors can be seen by examining flows between specific locations. For example, education majors who originate in New York are highly unlikely to work as teachers in New York unless they hold a master's degree. As a result, there is a large outflow of bachelor's-level education majors from New York to areas where working as a bachelor's-level teacher is more common, but where the wage returns to doing so are much lower. Migration flows such as these show that non-wage factors, specifically related occupation availability, are potentially strong determinants of the observed returns to college majors.

One would expect selection to result in naive estimates being upward biased if certain majors are more prone to migrate or choose a particular occupation in response to favorable wage shocks. On the other hand, naive estimates may be downward biased if certain majors have strong non-wage preferences for a particular location or occupation. Estimating the direction and magnitude of this bias is the primary empirical question of this paper.

To account for the various factors described above, I estimate an extended Roy (1951) model that allows for nonpecuniary tastes in both the location and occupation dimensions. The model divides occupations for each major into related and unrelated, and divides the United States into 15 groups of states. This paper bridges previous work that has examined the role of selective migration on the wage returns to a college degree (Dahl, 2002; Bayer, Khan, and Timmins, 2011) and the role of selective occupational choice on the returns to college major (Lemieux,

2014; Kinsler and Pavan, 2015).

Estimation of an extended Roy model is difficult in a model with nonpecuniary preferences and many choice alternatives. To estimate the model, I implement methods pioneered by Lee (1983) and Dahl (2002) which show that a control function approach, where the control function includes a polynomial of a small number of observed choice probabilities, is able to account for a variety of patterns in selection. 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 implement the Lee and Dahl approach with a machine learning method known as the conditional inference classification tree. While existing methods have utilized nonparametric bin estimation to derive selection probabilities, tree classification of this type has the advantage of using the data to determine which covariates should be included, and where bin cut points should be made. It also ensures that the selection probabilities are not overfit, meaning that the out-of-sample prediction remains good. The algorithm is especially useful in settings where it would be infeasible to include all covariates. I assess the performance of the classification tree relative to classical econometric estimators and show that it performs better both in simulation and in practice.

Using these empirical methods, I find that OLS estimates of the returns to college majors (relative to education majors) are upward biased. Correcting for selective migration and occupational choice tends to lower the measured returns, by up to 30% in some locations and consistent with other studies (Dahl, 2002; Bayer, Khan, and Timmins, 2011). The bias is the strongest among business and STEM majors who hold advanced degrees, as well as in locations in the Northeastern US. These results underscore the geographic specificity of the wage returns to major, as well as the overstatement of naive estimates of these returns.

2 A Roy Model of Migration, Occupation, and Earnings

In this section, I introduce an extended Roy (1951) model of college major, occupational choice, and locational choice, using the framework developed in Dahl (2002). It extends Roy's original model in two ways: (i) both pecuniary and nonpecuniary factors influence an individual's decision; and (ii) there are more than two alternatives in the choice set.

The focus of this paper is on how selective migration and occupational choice in the United States affect the measured returns to the human capital investment of college major. The objective is to examine how sensitive earnings in a particular major are to selectivity in post-college location and occupational choice. Existing models in the literature on college major and occupation have treated location as fixed (Lemieux, 2014; Kinsler and Pavan, 2015; Ransom and Phipps, 2017). At the same time, there is strong evidence that location is an increasingly important determinant of labor market outcomes, particularly for the college educated (Moretti, 2012; Diamond, 2016). This paper serves to fill the gap between these two literatures.

An extended Roy model serves as an appropriate lens through which to view the joint location and occupation decisions of college graduates because it allows for the inclusion of nonpecuniary components. Factors such as amenities and distance have been shown to be important determinants of migration decisions (Kennan and Walker, 2011; Koşar, Ransom, and van der Klaauw, 2020), while nonpecuniary considerations have also been shown to be important to occupational choice among college graduates (Keane and Wolpin, 1997; Arcidiacono et al., 2020).

I now formalize each component of the Roy model and how each of the components interact with each other. The primary components of the model are earnings (the outcome equation) and preferences (the selection equation). In contrast with

most of the Roy model literature, this paper emphasizes the empirical results of the outcome equation as opposed to the selection equation. As such, it is appropriate to view the model as a reduced-form approximation of a Roy model because I make no attempt to structurally model 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. groups of 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 nonpecuniary aspects of the chosen location and occupation. I discuss the reasons for and limitations of this two-period modeling assumption in Online Appendix C.

2.1 Earnings

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

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

where x_i is a vector of individual characteristics and s_i is an S-dimensional vector of dummy variables indicating i's college major and advanced degree attainment. Both x_i and s_i are assumed to be exogenous; see Online Appendix C for further discussion. The parameter of interest in (2.1) is $\gamma_{2\ell k}$, which measures the link between earnings, college major, and potential location and occupational choices. However, because $\eta_{i\ell k}$ is only observed in the chosen (ℓ, k) combination, and because the chosen (ℓ, k) is the result of a non-random selection process, OLS estimates of $\gamma_{1\ell k}$ and $\gamma_{2\ell k}$ will generally be biased.

2.2 Nonpecuniary utility

The nonpecuniary utility individual i receives from residing in location ℓ and working in occupation k given birth in location j is given by:

$$u_{ij\ell k} = z_i \alpha_{j\ell k} + \varepsilon_{ij\ell k}, \;\; \ell = 1, \ldots, L \;, \;\; k = 1, \ldots, K \tag{2.2}$$

where z_i is a vector of individual characteristics which may also include elements of x_i or s_i. $u_{ij\ell k}$ encompasses all nonpecuniary utility components that could determine the utility of residing in location ℓ and working in occupation k given origin j. These include location characteristics (climate, geography, distance from birth location, etc.) and occupational characteristics (working conditions, relevance to previous human capital investments, etc.).

I treat the parameters of this equation as nuisance parameters, since my objective is to uncover selectivity in earnings and not to measure the locational and occupational preferences of individuals.

2.3 Overall preferences

Individuals have preferences for both earnings and nonpecuniary factors:

$$V_{ij\ell k} = w_{i\ell k} + u_{ij\ell k}, \;\; \ell = 1, \ldots, L \;, \;\; k = 1, \ldots, K \tag{2.3}$$

The overall preferences can be rewritten in terms of conditional population means and individual-specific errors, as follows:

$$\begin{aligned} V_{ij\ell k} &= \underbrace{\mathbb{E}\left[w_{i\ell k} \mid x_i, s_i\right] + \mathbb{E}\left[u_{ij\ell k} \mid z_i\right]}_{v_{j\ell k}} + \underbrace{\eta_{i\ell k} + \varepsilon_{ij\ell k}}_{e_{ij\ell k}} \\ &= v_{j\ell k} + e_{ij\ell k} \end{aligned} \tag{2.4}$$

where $\eta_{i\ell k}$ represents individual unobserved determinants of earnings, and $\varepsilon_{ij\ell k}$

represents preference shocks for choosing to live in ℓ and work in occupation k given birth location j. $v_{j\ell k}$ is referred to as either the subutility function (in the selection literature) or the conditional value function (in the dynamic discrete choice literature). It is important to note that, being a conditional population mean, $v_{j\ell k}$ is fixed. It is also important to point out that the $(e_{ij\ell k})_{j,\ell,k}$ are assumed to be mean-independent of the regressors (x_i, s_i, z_i).

2.4 Utility maximization

Individuals maximize utility such that

$$d_{ij\ell k} = 1\left[v_{j\ell k} + e_{ij\ell k} \geq v_{jmn} + e_{ijmn} \ \forall\,(m,n) \neq (\ell,k)\right] \tag{2.5}$$

where $1[A]$ is an indicator variable that takes a value of 1 when condition A is true and 0 otherwise. (2.5) 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.

2.5 Selection rule

The selection rule is given by

$$w_{i\ell k} \text{ observed} \iff d_{ij\ell k} = 1 \tag{2.6}$$

Specifically, earnings are only observed if all L selection equations in (2.5) are simultaneously satisfied. Thus, individuals observed to reside in ℓ are not a random

sample of the population; hence

$$\begin{aligned}\mathbb{E}\left[\eta_{i\ell k} \,|\, w_{i\ell k} \text{ observed}, x_i, s_i, z_i\right] &= \mathbb{E}\left[\eta_{i\ell k} \,|\, d_{ij\ell k} = 1, x_i, s_i, z_i\right] \\ &= \mathbb{E}\left[\eta_{i\ell k} \,|\, e_{ijmn} - e_{ij\ell k} \leq v_{j\ell k} - v_{jmn},\ \forall (m,n) \neq (\ell,k)\right] \\ &\neq 0 \end{aligned} \tag{2.7}$$

where $\mathbb{E}\left[\eta_{i\ell k} \,|\, \cdot\right]$ is the selectivity bias for i, and where the second equality relies on the maintained assumption that the $(e_{ij\ell k})_{j,\ell,k}$ are mean-independent of the regressors (x_i, s_i, z_i).

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 nonpecuniary 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 nonpecuniary dimension, but which are unobserved or poorly measured.

In the next section, I explain how I avoid these issues by implementing existing estimation methods 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

In this section, I provide intuition and a brief formal derivation on how to feasibly estimate the aforementioned extended Roy model. I also informally discuss how the model is identified. The key point is that I express the selection in the earnings equation as a function of a small number of observed choice probabilities.

3.1 Overview

The intuition of this approach is as follows: examining equations (2.5) and (2.6) reveals that the probability of observing an individual's earnings in location ℓ and occupation k is related to the probability that $V_{ij\ell k}$ 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 ($e_{ij11} - e_{ijmn}, \ldots, e_{ijLK} - e_{ijmn}$) 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, referred to as *index sufficiency*, is that this bivariate distribution does not depend on the subutility functions themselves, except through a small number of choice probabilities. 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.

3.2 Technical details

To aid the exposition, I now briefly formalize the above intuition. Readers interested in a full derivation should consult Dahl (2002) and Lee (1983). Bourguignon, Fournier, and Gurgand (2007) is another excellent source which I follow below.

First consider a reformulation of (2.5) and (2.6):

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

where $\xi_{i\ell k} \equiv \max_{(m,n)\neq(\ell,k)} \left(v_{jmn} - v_{j\ell k} + e_{ijmn} - e_{ij\ell k}\right)$. Note that the zero on the second line is bolded to emphasize that it is a vector, but the zero on the final line is a scalar. Recall that the v's are conditional on the regressors (x_i, s_i, z_i) and that the e's are assumed to be mean-independent of the regressors.

I can now express the bias correction in terms of the probability that $\xi_{i\ell k} \leq 0$, the joint distribution of $\eta_{i\ell k}$ and $\xi_{i\ell k}$, and the covariates that enter the utility (z_i). Note that, since I do not make any parametric assumptions about the distributions of the $\eta_{i\ell k}$'s or $\varepsilon_{ij\ell k}$'s (and hence the $e_{ij\ell k}$'s), there is no closed-form expression for $\mathbb{P}(\xi_{i\ell k} \leq 0)$.

Let Γ_i be a vector containing a non-parametric function of the z_i's (which also includes x_i and s_i) across choice alternatives. Following Bourguignon, Fournier, and Gurgand (2007), I can express the selection bias as follows:

$$\begin{aligned} \mathbb{E}\left(\eta_{i\ell k} | \xi_{i\ell k} \leq 0, \Gamma_i\right) &= \iint_{-\infty}^{0} \frac{\eta_{i\ell k} f\left(\eta_{i\ell k}, \xi_{i\ell k} | \Gamma_i\right)}{\mathbb{P}\left(\xi_{i\ell k} \leq 0 | \Gamma_i\right)} d\xi_{i\ell k} d\eta_{i\ell k} \\ &= \mu\left(\Gamma_i\right) \end{aligned} \tag{3.2}$$

Assume that $\mu(\cdot)$ is invertible and denote $p_{ij\ell k}$ as the probability that i prefers (ℓ, k) given birth in location j. Then (3.2) can be equivalently written as

$$\mathbb{E}\left(\eta_{i\ell k} | \xi_{i\ell k} \leq 0, \Gamma_i\right) = \lambda\left(p_{ij11}, \ldots, p_{ijLK}\right) \tag{3.3}$$

In order to simplify estimation, Dahl proposes an *index sufficiency assumption* as follows:

$$\begin{aligned} f\left(\eta_{i\ell k}, \xi_{i\ell k} | \Gamma_i\right) &= f\left(\eta_{i\ell k}, \xi_{i\ell k} | p_{ij11}, \ldots p_{ijLK}\right) \\ &= f\left(\eta_{i\ell k}, \xi_{i\ell k} | p_{ij\ell k}, p_{ijmn}\right) \end{aligned} \tag{3.4}$$

where $p_{ij\ell k}$ and p_{ijmn} are two probabilities that are readily observed in the data. In general, index sufficiency refers to any number of probabilities that are a strict subset of the entire set of probabilities. For simplicity, and in the spirit of Dahl (2002), I assume two. Stated differently, the assumption in (3.4) is that the probabilities $p_{ij\ell k}$ and p_{ijmn} contain all of the relevant information in Γ_i.

Applying the assumption in (3.4) to the earnings equation gives the following set of corrected earnings equations that account for selective migration and occupational choice, and that are feasibly estimated:

$$w_{i\ell k} = x_i \gamma_{1\ell k} + s_i \gamma_{2\ell k} + \sum_{j=1}^{L} d_{ij\ell k} \lambda_{j\ell k}\left(p_{ij\ell k}, p_{ijmn}\right) + \omega_{i\ell k}, \tag{3.5}$$

The implication of the assumption in (3.4) is that $\mathbb{E}\left[\omega_{i\ell k} \mid x_i, s_i, p_{ij\ell k}, p_{ijmn}, d_{ij\ell k} = 1\right] = 0$, meaning that the selection problem has been resolved.

Because index sufficiency is an assumption, it is important to recognize the restrictions that it imposes. Index sufficiency holds, for example, if earnings errors are composed of an individual fixed effect that is invariant to the location of residence (see Dahl (2002) for further details). On the other hand, this assumption is less likely to hold in a setting where, for example, an individual's fixed effect on earnings could vary with location. Another example of an index sufficiency violation would be an individual-specific location match quality as specified in Kennan and Walker (2011). I discuss in Online Appendix A the results of Monte Carlo simulations that show that this assumption holds for a variety of scenarios.

In estimation, I make two additional simplifying assumptions in (3.5). First, I

assume that the selection correction functions are the same for everyone, i.e. that the correction term can be rewritten as $\sum_j \lambda_{j\ell k}(\cdot) = \lambda_{\ell k}\left(p_{ij\ell k}, p_{ijmn}\right)$. While this assumption is restrictive, it allows me to estimate the wage effect of staying in the birth location. Second, I assume that the unknown correction functions $\lambda_{\ell k}(\cdot)$ can be well approximated by a fully interacted cubic polynomial in $\left(p_{ij\ell k}, p_{ijmn}\right)$ (Dahl, 2002).

For the choice of probabilities $p_{ij\ell k}, p_{ijmn}$, I assign $p_{ij\ell k}$ to be the first-best choice probability, and p_{ijmn} to be the probability that individual i would live 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.

3.3 Identification

I now informally discuss how the model is identified. As discussed in other implementations of the Roy model (Dahl, 2002; Bayer, Khan, and Timmins, 2011; D'Haultfœuille and Maurel, 2013), separately identifying nonpecuniary preferences from earnings in most cases requires an exclusion restriction—a covariate which appears in the choice probabilities but does not affect wages (see D'Haultfœuille and Maurel (2013) for an exception).

Crucial to identification in this model is the existence of two such exclusion restrictions: one for locational choice and one for occupational choice. I use two related exclusion restrictions inspired by Kinsler and Pavan (2015). To separately identify preferences for location from earnings, I use the fraction of demographically similar (including college major and advanced degree status) individuals from the same birth state who stayed in their birth state, net of the national rate of staying. To separately identify preferences for occupation from earnings, I compute a similar number, but instead calculate the share who choose to work in an occupation related to their major.

The intuition for these exclusion restrictions is as follows: a person who ends

up leaving a "sticky" state (in which most similar-looking people stay) must have a low preference for living there (or, equivalently, be more responsive to earnings differences); and vice versa for someone who ends up staying in a state where most similar-looking people leave. Similar logic applies to the related occupation decision.

The ideal exclusion restriction for location or occupational choice would be an adequate measure of search frictions. This is because the observed location or occupational choice is a result of labor supply and labor demand factors. Being able to isolate labor demand through search frictions would allow me to tell whether someone's chosen location or occupation is due to utility maximization, or because they could not get a job in the preferred location or occupation.[3] While not a perfect measure of search frictions, the proposed exclusion restrictions recover a reduced-form approximation of such.

The peer share exclusion restrictions described above would be invalidated if these shares had a direct effect on wages. This could arise as a result of general equilibrium factors and could be particularly problematic for the occupational choice. For example, if there is an excess supply of stayers in a major, this could directly drive down the corresponding related-occupation wage in the home location.

An advantage of using the above exclusion restrictions is that it allows me to include birth location directly in the wage equation. Previous literature has shown that certain locations do a better job of educating their residents, which implies that stayers in those locations may receive higher wages than movers (Card and Krueger, 1992; Heckman, Layne-Farrar, and Todd, 1996; McHenry, 2011). Allowing stayers to earn different wages than movers improves on the previous approaches of Dahl (2002) and Bayer, Khan, and Timmins (2011) which each require birth location to be excluded from wages.

In addition to the peer share exclusion restrictions, I also allow distance moved

[3]This reasoning presumes that individuals would want to have a job in hand before moving to a new location. Among college graduates, this presumption is likely to be correct (see Balgova, 2018).

and other demographic characteristics to influence the nonpecuniary portion of utility but not earnings. Specifically, these covariates are: an indicator for birth location in the same Census region as the location of residence, and separate indicators for each of the following: co-residence with a family member, spouse's work status (if applicable), spouse born in residence location, and presence of children aged 0-4 or 5-18. In results not shown, but available upon request, I find that these demographic characteristics have much less predictive power in the first stage (i.e. the location choice) than the two primary exclusion restrictions.

4 Data and Descriptive Analysis

I now discuss the data used in the estimation procedure. 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. (2020) over the years 2010–2019. The analysis sample consists of all native-born males between the ages of 22 and 54 with at least a bachelor's degree, and who report earnings within a reasonable range, who report their 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 10% sample of the US population for this subgroup. The estimation sample of the data comprises 1.024 million individuals. See Online Appendix B for more details about the ACS and Table F1 for sample selection criteria.

I now discuss aggregation of majors, occupations, and locations in order to preserve tractability in estimation.

Majors I aggregate majors from a set of 51 detailed majors into five broad categories, crossed with advanced degree status so that s_i in equation (2.1) is a 10-dimensional vector. The set of aggregated majors is: education, social sciences, business, STEM, and all others. Notably, the business field includes economics majors and the STEM field includes pre-med majors. See Online Appendix B for complete details.

Occupations I define occupation as having two values: *related* or *unrelated* (i.e. $K = 2$). An occupation is 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).[4] 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. I allow the set of related occupations to differ based on advanced degree status. See Online Appendix B for complete details.

Locations Because the empirical method employed in this paper does not work well in small samples, I aggregate locations as another way of maintaining statistical power. Specifically, I divide the United States into 15 locations, corresponding to states or groups of adjacent states. The 15 locations consist of the five largest states (California, Texas, Florida, New York, and Illinois), followed by the nine Census divisions, with the South Atlantic division being divided in two. A detailed list of each location is included in Table F7. Aggregating locations in this way loses only 14% of inter-state moves. Thus, it is unlikely to meaningfully bias my estimates.

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

4.2 Descriptive Analysis

To motivate the modeling approach described in Section 2, I now discuss descriptive evidence of the heterogeneity of migration and occupational choice across majors at the national level, and heterogeneity in migration flows across certain locations by college major, advanced degree status, and occupation.

4.2.1 Summary statistics

Table 1 lists differences across major in the three outcomes considered in this paper. The results in the odd-numbered rows of the table are regression coefficients on major dummies, estimated at the national level and controlling for demographics, advanced degree status, CBSA fixed effects, and a cubic in potential experience. The results in parentheses are standard deviations of the distribution of state-level coefficients.

The results of Table 1 show that education majors earn the least, leave their birth state at the lowest rates, and work in related occupations at the highest rates. What is interesting from the table is that there is no clear monotonicity among these three outcomes. For example, STEM and business majors each earn about the same amount and work in related occupations at similar rates. However, STEM majors are much more likely to leave their state of birth.

Finally, the standard deviations in Table 1 show that there is substantial heterogeneity in these outcomes across states, and that cross-state variation in migration and availability of related occupations is as large as cross-state variation in earnings. While the spatial variation in earnings is well known, variation in migration and concentration of related occupations is much less known. As discussed previously, variation in these latter two outcomes is a crucial component of identification of the extended Roy model.

4.2.2 Transition Matrix

The results of the previous subsection indicate that there is sizable variation across locations in all three of the outcomes that I consider. In this section, I present evidence on how migration flows are related to the variation in location-specific outcomes previously documented.

Figure 1 displays the migration transition matrix by major for the five largest states, for those who do not hold an advanced degree. Rows indicate birth location, while columns indicate residence location. Each row and column contains five bars, which correspond to the five majors. Each bar is divided in two, with the bottom section corresponding to the share of individuals choosing the related occupation.

Examining Figure 1 reveals a number of examples that support the model. First, the flow of workers from New York to Florida is remarkable. Underscoring this pattern is the fact that Florida is disproportionately popular for New York education majors. Furthermore, it is especially evident of non-pecuniary factors because the education majors who stay in New York disproportionately leave the teaching occupation, while those who move to Florida are disproportionately in the education occupation. The reverse is also true: education majors who leave Florida (see the second row) are almost all those who choose the non-education occupation.

Figure 2 is the transition matrix for advanced degree holders. While there are high flows from New York to Florida among this group, there are equally high flows from New York to California. Furthermore, the education majors in New York who earn master's degrees stay in New York and work as teachers at much higher rates than their counterparts who do not hold master's degrees. These findings are further evidence of self-selection in location and occupation decisions that differ by college major and advanced degree status.

It is worth noting one other observation from Figures 1 and 2. Examining the middle bar of the off-diagonal elements of columns 1 and 4 shows the fraction of

other majors who choose to move to California and New York. Of the movers who choose these two locations, other majors are disproportionately represented. This likely reflects the fact that other majors are composed of performing arts majors, and California and New York are hubs for such occupations. This is consistent with migration being a function not only of earnings, but also of availability of related occupations.

5 Estimation

In this section, I discuss how to estimate the final equation (3.5) of the model discussed in Sections 2 and 3. The estimation proceeds in two stages. First, I estimate the choice probabilities $\left(p_{ij\ell k}, p_{ijmn}\right)$. Second, I estimate the parameters of equation (3.5), including the cubic-approximated correction functions $\lambda_{\ell k}$.

5.1 Choice probabilities

There are a variety of ways to estimate the choice probabilities. The most popular are the conditional logit model and non-parametric estimation.

The conditional logit model is by far the most popular method used to estimate choice probabilities (and in migration models in particular, because the dimension of the choice set tends to be 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 (IIA) property.

Non-parametric estimation has two advantages. First, it does not require the researcher to model location-specific characteristics, of which there are a large number and of which many are poorly measured. Second, it does not require the researcher to specify the dependence structure of the choice alternatives for a non-IIA model.

The primary drawback to non-parametric estimation is deciding how finely

and in which ways to divide the state space. For tractability, probabilities must be discretized, and probabilities that are estimated from cells that are too small will 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, Hornik, and Zeileis (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 aggregating 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 and in Online Appendix E.[5]

In statistical applications, machine learning amounts to using methods that combine estimation with model selection to enhance out-of-sample prediction of statistical models. The result is an algorithm which automatically selects which covariates to include while also estimating their parameters. In the current setting, the conditional inference recursive partitioning algorithm selects which variables and which levels of the variables matter most in predicting migration and occupations.

[5]There are other non-parametric machine learning methods. For example, Snoddy (2019) uses a random forest to estimate a similar model as mine.

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 result, poorly fitting out of sample) such that out-of-sample prediction is maximized.[6] The algorithm works for both continuous and categorical variables on both sides of the equation.[7] The current application contains a categorical dependent variable and covariates that are primarily categorical, but some of which are continuous.

I detail the algorithm in Online Appendix E. As a short summary, the algorithm iterates on selecting covariates and then splitting those covariates to produce terminal nodes that are as "pure" as possible, i.e. nodes that predict as much as possible one category of the dependent variable. Stopping is determined by whether additional covariates significantly affect prediction conditional on the nodes already created.

As an example of what the output of this algorithm looks like, I include Figure 3, which depicts a stylized 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 output shows that the most distinct difference occurs when splitting at a value of three years of experience, followed by an additional split that occurs at a value of eight. At

[6] Hothorn, Hornik, and Zeileis (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.

[7] 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. The algorithm can also be used in survival analysis.

these splits, New York is entirely composed of individuals with less than four years of work experience, Texas is composed nearly perfectly of individuals with experience levels between four and eight years, and workers with nine or more years of experience almost certainly live elsewhere. In the actual estimation, each tree will have hundreds of terminal nodes (cells).

5.1.3 Implementation of the non-parametric estimation algorithm

I now discuss in detail the estimation of the choice probabilities and which variables are used to predict migration and occupational choice. The non-parametric nature of the probability estimation results in probabilities that are discretized into cells. So long as the instruments are included in the discretization process (analogously, so long as the instruments have a strong first stage), then the probabilities can be used in estimation as described earlier. One potential advantage of the tree estimator relative to a bin estimator is that the instrument may more flexibly enter the first stage, thus increasing the strength of the first stage (Angrist and Frandsen, 2019).

The conditional inference tree algorithm assigns cells based on the following characteristics: whether the individual was born in the location of residence or in the same Census region; college major; advanced degree status; age; race; marital status; 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 two exclusion restrictions discussed in Section 3.3. 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 choosing the alternative under consideration vs. all others. The estimated probabilities are then grouped according to the terminal nodes of the tree.

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 the traditional non-parametric bin estimator or with the logit estimator, the latter of which is by far the most popular estimation method for discrete choice models.

The primary benefits of the tree algorithm are twofold: (*i*) it allows the researcher to consider a large number of candidate covariates without having to worry about encountering the curse of dimensionality (i.e. the result of which would be empty bins); and (*ii*) it allows the sample space to be divided into irregularly shaped bins. The first benefit arises out of model selection and could be accomplished with other parameter regularization methods such as LASSO (Belloni, Chernozhukov, and Hansen, 2011). The second benefit arises out of the algorithm's recursive nature: by not making all splits simultaneously, the division of the state space can contain non-rectangular shapes. A final benefit of the algorithm is that it performs slightly better at out-of-sample prediction than existing methods. A summary of this is given in Table F8.

The benefits of the tree algorithm are manifest in Online Appendix A where I compare the small- and large-sample performance of various algorithms and error structures. The tree algorithm performs about as well as the bin estimator in large samples, but much better in small samples. Furthermore, if the researcher overly aggregates the bins (because of the curse of dimensionality), then the tree algorithm significantly outperforms the simple bin estimator.

While the tree algorithm performs better in Monte Carlo simulations, does it substantially alter the estimates of selection bias in the ACS data? The answer is yes. In results not shown, but available from the author upon request, I find that using either a bin or a logit estimator causes the degree of selection bias to be understated. That is, the model estimates when using these two estimators tend to fall in between those of OLS and the tree algorithm, as in the simulation in Online

Appendix A. This evidence is further support for the appropriateness of the tree algorithm in this particular application.

5.2 Earnings equation

The earnings equation parameters in (2.1) are estimated by OLS (separate equations for each location and occupation) after making use of the index sufficiency assumption in (3.4) and the dimensionality reduction assumptions discussed previously (i.e. using a cubic polynomial of the two selection probabilities). I obtain standard errors by bootstrapping with 500 replications. In each bootstrap replication, I resample the data, estimate the selection probabilities using the tree algorithm previously discussed, and then estimate the earnings parameters using the sampled data and the estimated selection probabilities. Bootstrapping ensures that the standard errors of the earnings equation estimates appropriately account for sampling error in the estimated probabilities. Using an automated model selection algorithm such as the tree algorithm described above may pose additional challenges to inference by introducing specification error (as well as sampling error) into the earnings estimates (Mullainathan and Spiess, 2017). However, I leave to future research a more complete investigation of these issues.

6 Empirical Results

I now discuss the parameter estimates of the earnings equation with and without selection correction. For estimation results associated with the estimation of the choice probabilities, see Online Appendix E. The primary parameters of interest are the coefficients on the college major dummies and their interaction with a dummy for advanced degree attainment. The primary research question is how these parameter estimates change once I account for self-selection into locations and occupations. Throughout, I treat bachelor's-level education majors as the reference

category.

6.1 Estimates for specific states

Table 2 lists a subset of parameter estimates of equation (3.5) with the implemented simplifications discussed in Section 3.2. While I estimate 30 equations, I present detailed results for only three of the five most populous states. I later present aggregate results for all 15 locations.

Table 2 reports the earnings equation estimates for each occupation in the three states, for both the naive case and the corrected case. The first column within each state and occupation reports the estimated returns to each major assuming exogeneity and no selection bias, while the second column reports the estimated returns after correcting for selection into residence location and occupation but not selection into majors. The OLS estimate is upward biased for the majority of all coefficients. The magnitude of the upward bias differs from state to state, with the largest differences in New York and the smallest differences in Florida.

As noted previously, I am able to separately identify the earnings effect of stayers. These estimates are reported on the last row of Table 2. Without selection correction, there appears to be a wage penalty for stayers in each of these three states. However, this penalty gets erased once correcting for selection, indicating that what naively appears to be a compensating differential for staying in one's birth state is actually a selection effect.

It is important to keep in mind the interpretation of what generates the direction of the bias in returns. As noted in Dahl (2002) and Bayer, Khan, and Timmins (2011), an upward bias in the returns to schooling is the result of individuals responding to above-average earnings shocks. This comes about in the model through the selection correction terms: if someone moves to a location when observationally similar individuals do not, then it must be because of a favorable earnings shock. Put differently, moves in response to favorable earnings shocks will overstate the

treatment effect of college major or occupation, compared to randomly assigning individuals to live in a given location.

6.2 Estimates for all locations

A remaining question is whether or not the differences between the corrected and uncorrected estimates are statistically or economically significant. I test for this in two ways: (i) I conduct an F test for joint significance of the polynomial correction terms; and (ii) I conduct a Hausman-type test where the null hypothesis is that the baseline OLS is efficient and consistent, while the corrected estimates are consistent but inefficient. I present a more complete discussion of this process in Online Appendix D.

I consider an estimate to be statistically and economically significant if the following criteria hold: (i) the Hausman test statistic overturns the null hypothesis of no difference between corrected and uncorrected at the 5% level; (ii) both the uncorrected and corrected coefficients are statistically different from zero at the 5% level; (iii) the percentage difference between the corrected and uncorrected returns is significantly different from zero at the 5% level; and (iv) the percentage difference exceeds 10% in magnitude. I obtain standard errors of the percentage change in returns from the bootstrapping approach described previously.

Table 3 lists moments of the distribution of the percentage change in returns when correcting for selection, as well as how many estimates are significantly different, using the definition above. At the median, selection does not significantly change the measured returns for any majors except those with advanced degrees who work in a related occupation. Among this group, the Business and STEM fields have the largest number of significantly different estimates. The median percentage change corresponds to an upward bias of about 15%. Comparison within each set of estimates reveals that locations in the Northeastern US—New York, New Jersey/Pennsylvania, and New England—usually have the largest magnitude of

bias.[8] This suggests that there is a spatial component to the selection bias.

7 Conclusion

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

To estimate the model, I implement the framework of Dahl (2002) and Lee (1983) which allows 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 on college-educated men from the American Community Survey from years 2010–2019. I also illustrate the advantages of using machine learning methods to non-parametrically estimate the selection probabilities.

I find that selective migration and occupational choice cause an upward bias in the measured wage returns to college major, relative to education majors. The percentage change in the corrected returns ranges from 0% to 28% for STEM and business majors, is strongest among advanced degree holders, and is statistically and economically significant in about one-third of all locations.

The results underscore the importance of appropriately measuring returns to human capital investment. It is well established that students choose majors in part because of earnings differences. What is less clear is if students know that their earnings outcomes may be geographically specific. Students' beliefs on spatial differences in earnings also motivate future research to consider how location preferences and spatial earnings differences feed back into major choices (Wiswall and Zafar, 2015; Arcidiacono et al., 2020).

[8] Results on the precise values of the returns to major for all locations for each of the two occupations and advanced degree statuses are shown in Tables F10–F27. Figures F3–F6 plot uncorrected and corrected returns for all majors, locations, and occupations.

References

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2016. "Cashier or Consultant? Entry Labor Market Conditions, Field of Study, and Career Success." *Journal of Labor Economics* 34 (S1):S361–S401.

Angrist, Joshua and Brigham Frandsen. 2019. "Machine Labor." Working Paper 26584, National Bureau of Economic Research.

Arcidiacono, Peter, V. Joseph Hotz, Arnaud Maurel, and Teresa Romano. 2020. "*Ex Ante* Returns and Occupational Choice." *Journal of Political Economy* 128 (12):4475–4522.

Balgova, Maria. 2018. "Why Don't Less Educated Workers Move? The Role of Job Search in Migration Decisions." Working paper, Oxford University.

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 (2):201–215.

Beffy, Magali, Denis Fougère, and Arnaud Maurel. 2012. "Choosing the Field of Study in Postsecondary Education: Do Expected Earnings Matter?" *Review of Economics and Statistics* 94 (1):334–347.

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.

Bourguignon, François, Martin Fournier, and Marc Gurgand. 2007. "Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons." *Journal of Economic Surveys* 21 (1):174–205.

Card, David and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 100 (1):1–40.

Dahl, Gordon B. 2002. "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets." *Econometrica* 70 (6):2367–2420.

D'Haultfœuille, Xavier and Arnaud Maurel. 2013. "Inference on an Extended Roy Model, with an Application to Schooling Decisions in France." *Journal of Econometrics* 174 (2):95–106.

Diamond, Rebecca. 2016. "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000." *American Economic Review* 106 (3):479–524.

Heckman, James, Anne Layne-Farrar, and Petra Todd. 1996. "Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings." *Review of Economics and Statistics* 78 (4):562–610.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1):153–161.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. 2006. "Unbiased Recursive Partitioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics* 15 (3):651–674.

Hothorn, Torsten and Achim Zeileis. 2015. "partykit: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research* 16 (12):3905–3909.

Keane, Michael P. and Kenneth I. Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy* 105 (3):473–522.

Kennan, John and James R. Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions." *Econometrica* 79 (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 (4):933–972.

Koşar, Gizem, Tyler Ransom, and Wilbert van der Klaauw. 2020. "Understanding Migration Aversion Using Elicited Counterfactual Choice Probabilities." Working Paper 8117, CESifo. URL https://ssrn.com/abstract=3544503.

Lee, Lung-Fei. 1983. "Generalized Econometric Models with Selectivity." *Econometrica* 51 (2):507–512.

Lemieux, Thomas. 2014. "Occupations, Fields of Study and Returns to Education." *Canadian Journal of Economics* 47 (4):1047–1077.

McHenry, Peter. 2011. "The Effect of School Inputs on Labor Market Returns that Account for Selective Migration." *Economics of Education Review* 30 (1):39–54.

Moretti, Enrico. 2012. *The New Geography of Jobs*. New York: Houghton Mifflin Harcourt.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2):87–106.

Ransom, Michael R and Aaron Phipps. 2017. "The Changing Occupational Distribution by College Major." *Research in Labor Economics* 45 (1).

Roy, A.D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3 (2):135–146.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek. 2020. *IPUMS USA: Version 10.0 [dataset]*. Minneapolis, MN: IPUMS.

Snoddy, Iain. 2019. *Essays in the Economics of Local Labour Markets*. Ph.D. thesis, University of British Columbia.

Winters, John V. 2017. "Do Earnings by College Major Affect College Graduate Migration?" *Annals of Regional Science* 59:629–649.

Wiswall, Matthew and Basit Zafar. 2015. "Determinants of College Major Choice: Identification using an Information Experiment." *Review of Economic Studies* 82 (2):791–824.

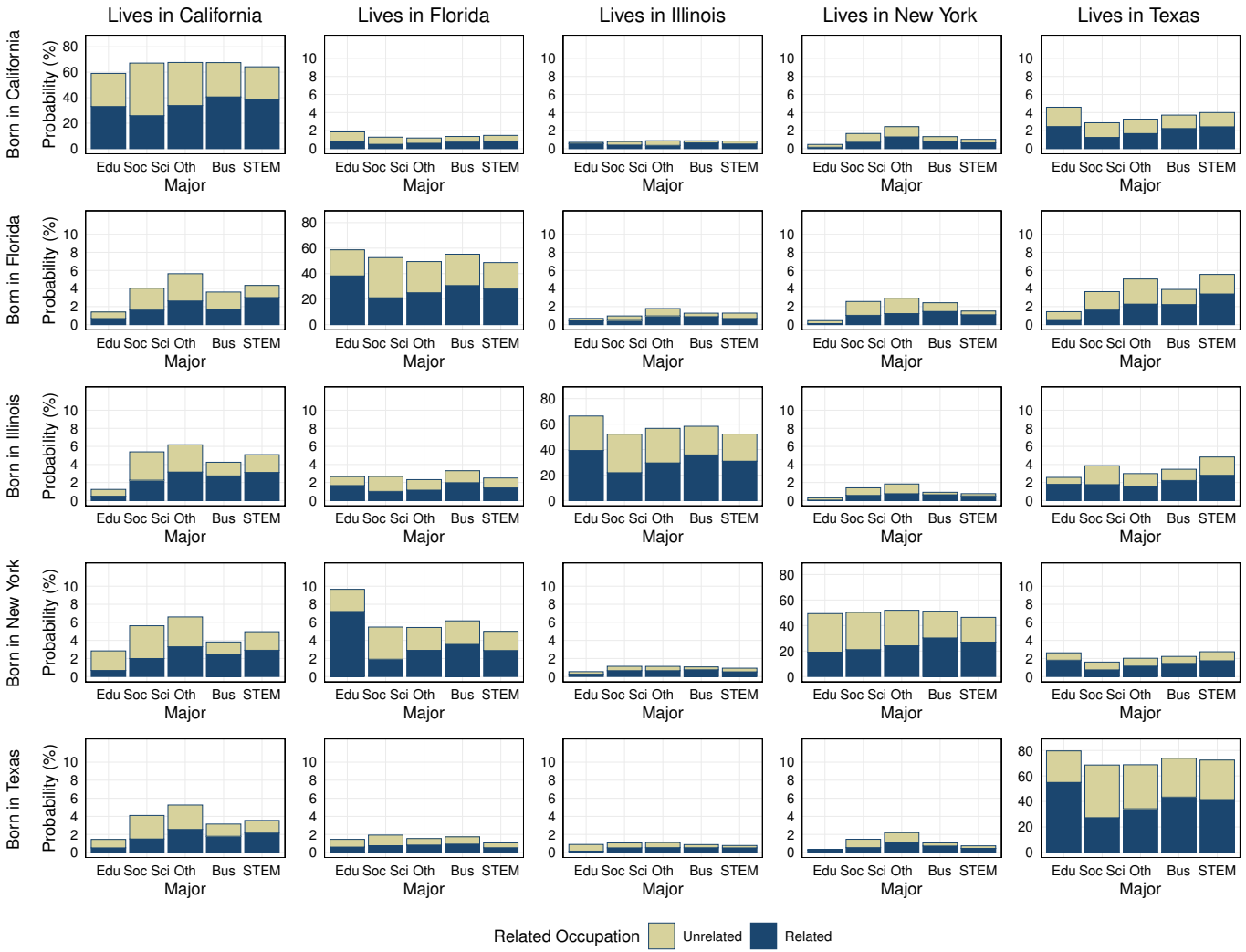
Figures and Tables

Table 1: Differences in outcomes by college major, relative to education majors

	Education	Soc Sci	Other	Business	STEM
Log Earnings	0.00	0.192	0.165	0.411	0.425
	(—)	(0.062)	(0.053)	(0.06)	(0.06)
Pr(Lives outside birth state)	0.00	0.118	0.126	0.08	0.134
	(—)	(0.059)	(0.063)	(0.056)	(0.062)
Pr(Works in related occupation)	0.00	-0.167	-0.115	-0.027	-0.03
	(—)	(0.04)	(0.044)	(0.049)	(0.054)
Frequency (%)	5.29	11.49	21.09	27.35	34.77
N	54,171	117,593	215,872	279,971	355,918

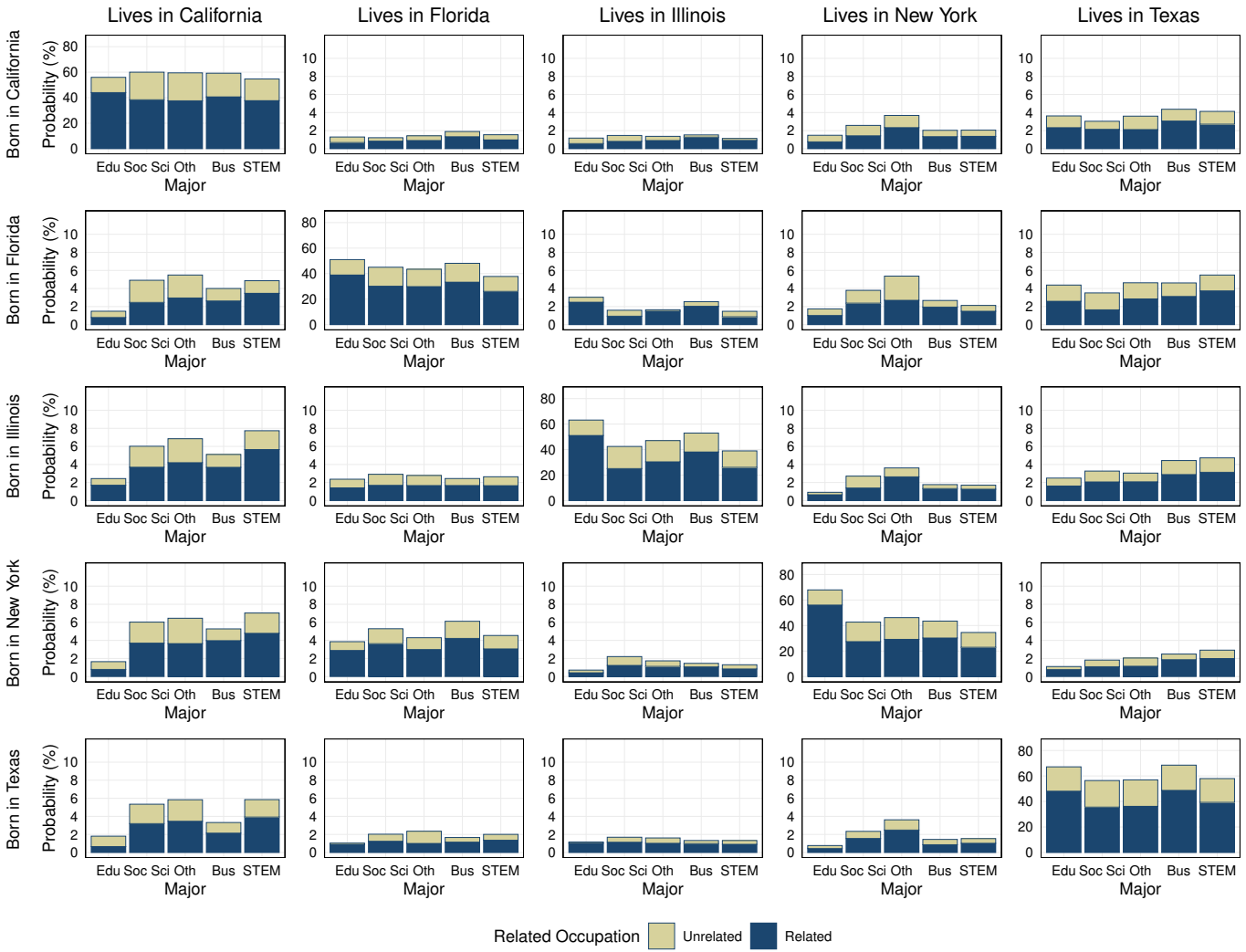
Notes: Regression estimates at national level, controlling for demographics, advanced degree status, CBSA dummies, and a cubic in potential experience. Standard deviation of state-specific estimates reported below in parentheses. All variables except for log earnings and distance are expressed in percentage points and estimated from linear probability models. Sample taken from the 2010-2019 American Community Survey and is restricted to males ages 22-54 with a bachelor's degree or higher. Sample weights are included in the computation. Additional details on sample selection can be found in Table F1.

Figure 1: Migration and occupation transition matrix by major for the five largest states: Non-adv. deg. holders



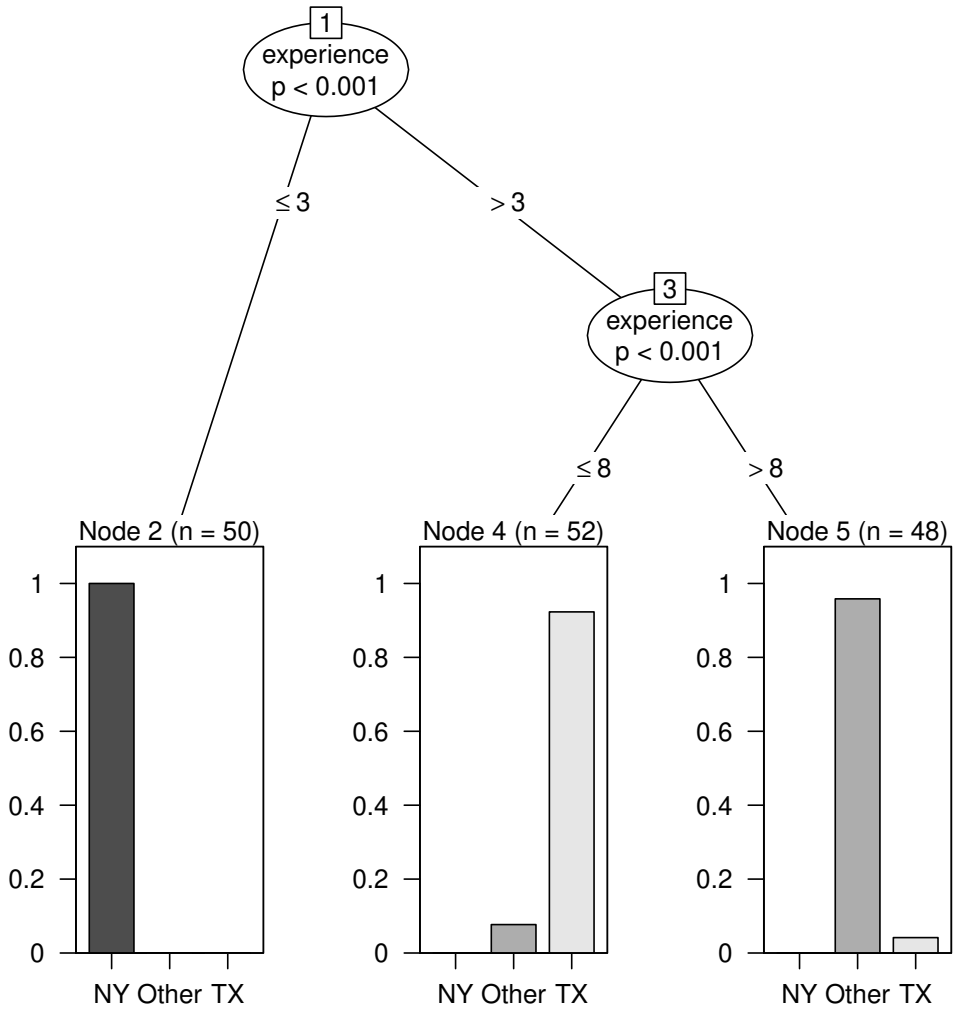
Notes: Markov transition matrix probabilities of living in a particular location and working in a particular occupation, by major, for the five largest US states. Light-colored bar segments represent proportion working in an unrelated occupation. Dark-colored bar segments represent proportion working in a related occupation.

Figure 2: Migration and occupation transition matrix by major for the five largest states: Adv. degree holders



Notes: Markov transition matrix probabilities of living in a particular location and working in a particular occupation, by major, for the five largest US states. Light-colored bar segments represent proportion working in an unrelated occupation. Dark-colored bar segments represent proportion working in a related occupation.

Figure 3: Simple example of tree structure from conditional inference recursive partitioning algorithm



Notes: Sample tree output from fictitious data using the algorithm described in Section 5.1.2. In this example, there are three locations to choose from: New York (NY), Texas (TX), or elsewhere (Other). The bars in Nodes 2, 4 and 5 represent the probability of choosing each location conditional on the characteristics of each node (i.e. experience ≤ 3, experience $\in (3, 8]$, or experience > 8, respectively).

Table 2: Uncorrected vs. corrected earnings equation estimates for select states

	Florida				New York				Texas			
	Unrelated Occupation		Related Occupation		Unrelated Occupation		Related Occupation		Unrelated Occupation		Related Occupation	
	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected	Uncorrected	Corrected
Bachelor's degree												
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)	(0.000)	(0.000)	(0.000)	(0.000)
Social sciences major	0.060*	0.062*	0.365***	0.377***	0.105***	0.071**	0.241***	0.222***	0.054*	0.053	0.192***	0.218***
	(0.033)	(0.034)	(0.023)	(0.024)	(0.030)	(0.032)	(0.033)	(0.036)	(0.031)	(0.033)	(0.023)	(0.026)
Other major	0.075**	0.072**	0.309***	0.316***	0.029	0.005	0.184***	0.161***	0.059**	0.048	0.167***	0.179***
	(0.031)	(0.032)	(0.017)	(0.019)	(0.027)	(0.030)	(0.030)	(0.032)	(0.029)	(0.030)	(0.016)	(0.018)
Business major	0.202***	0.195***	0.537***	0.533***	0.218***	0.158***	0.478***	0.423***	0.195***	0.190***	0.430***	0.429***
	(0.030)	(0.031)	(0.016)	(0.016)	(0.027)	(0.031)	(0.029)	(0.032)	(0.028)	(0.029)	(0.015)	(0.016)
STEM major	0.205***	0.195***	0.492***	0.488***	0.203***	0.138***	0.416***	0.358***	0.275***	0.268***	0.385***	0.383***
	(0.031)	(0.031)	(0.016)	(0.016)	(0.027)	(0.032)	(0.028)	(0.032)	(0.028)	(0.029)	(0.015)	(0.016)
Advanced degree (interaction)												
Education major	0.185**	0.177**	0.053	0.037	0.193***	0.079	0.209***	0.125**	0.034	0.023	-0.145***	-0.168***
	(0.077)	(0.077)	(0.047)	(0.049)	(0.058)	(0.071)	(0.041)	(0.051)	(0.061)	(0.061)	(0.039)	(0.039)
Social sciences major	0.260***	0.245***	0.112**	0.086*	0.272***	0.190***	0.215***	0.146***	0.094*	0.079	0.031	-0.028
	(0.066)	(0.066)	(0.046)	(0.048)	(0.049)	(0.052)	(0.039)	(0.041)	(0.051)	(0.052)	(0.043)	(0.046)
Other major	0.167**	0.155**	0.102**	0.075	0.230***	0.145***	0.147***	0.090**	0.014	0.009	-0.024	-0.068*
	(0.065)	(0.065)	(0.046)	(0.047)	(0.043)	(0.046)	(0.033)	(0.037)	(0.049)	(0.049)	(0.036)	(0.037)
Business major	0.150**	0.140**	0.139***	0.125***	0.260***	0.207***	0.248***	0.205***	0.056	0.057	0.068**	0.042
	(0.063)	(0.063)	(0.041)	(0.042)	(0.048)	(0.049)	(0.031)	(0.034)	(0.049)	(0.049)	(0.034)	(0.035)
STEM major	0.231***	0.221***	0.269***	0.251***	0.350***	0.291***	0.183***	0.132***	0.086*	0.082*	0.123***	0.093***
	(0.060)	(0.059)	(0.041)	(0.042)	(0.045)	(0.047)	(0.030)	(0.032)	(0.046)	(0.047)	(0.035)	(0.035)
Born here	-0.044***	-0.016	-0.038***	0.016	-0.112***	-0.001	-0.118***	0.011	-0.081***	-0.025	-0.073***	0.018
	(0.011)	(0.024)	(0.010)	(0.019)	(0.011)	(0.024)	(0.008)	(0.017)	(0.009)	(0.019)	(0.006)	(0.014)
Cubic in experience	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Demographics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CBSA fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
F test for λ terms		6.034		9.504		29.037		47.313		7.244		18.653
		[0.000]		[0.000]		[0.000]		[0.000]		[0.000]		[0.000]
R^2	0.210	0.212	0.270	0.272	0.285	0.291	0.291	0.297	0.257	0.258	0.300	0.303
Observations	18,992	18,992	27,888	27,888	25,928	25,928	37,908	37,908	30,507	30,507	46,716	46,716

Notes: Cubic in experience is fully interacted with advanced degree status. The wage return for an advanced degree holder is the sum of the bachelor's degree coefficient and the advanced degree interaction. Bootstrapped standard errors (500 replicates) 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 3: Percentage change in returns when correcting for selection

	Unrelated occupation				Related occupation			
Major	p10	Median	p90	No. significant	p10	Median	p90	No. significant
Bachelor's degrees								
Education	0.0	0.0	0.0	0	0.0	0.0	0.0	0
Social Science	-39.7	-3.0	34.6	1	-4.3	2.0	9.0	1
Other	-25.7	-4.8	80.5	0	-8.8	0.5	6.7	2
Business	-12.1	-5.0	-1.7	4	-6.5	-0.8	0.4	1
STEM	-14.4	-4.6	-2.2	4	-8.8	-1.8	0.0	2
Advanced degrees								
Education	-60.1	-5.3	2.3	2	-98.3	-23.1	43.1	3
Social Science	-29.9	-10.1	-3.3	6	-61.1	-31.5	6.3	4
Other	-32.3	-10.2	-1.1	4	-116.6	-36.1	133.4	4
Business	-24.2	-7.0	-0.1	4	-28.1	-16.4	-5.0	6
STEM	-20.8	-5.6	-2.4	4	-28.1	-14.8	-4.9	6

Notes: Summary statistics of the 15-location distribution of the percentage change between uncorrected and corrected returns to majors. "No. significant" counts the number of locations satisfying the following conditions: (i) the Hausman test statistic overturns the null hypothesis of no difference between corrected and uncorrected at the 5% level; (ii) both the uncorrected and corrected coefficients are statistically different from zero at the 5% level; (iii) the percentage difference between the corrected and uncorrected returns is significantly different from zero at the 5% level; and (iv) the percentage difference exceeds 10% in magnitude. Percentage changes are least informative for education, social science, and other majors because these majors have bases (i.e. uncorrected returns) that may be very close to zero.

A Online Appendix: Monte Carlo Simulation

In this section I detail the Monte Carlo simulation used to compare the performance of the conditional inference tree estimator with more traditional estimators.

A.1 Data generating process

Consider the following data generating process, structured similarly to the model in Section 2.

$$w_{i\ell k} = x_i\gamma_{1\ell k} + s_i\gamma_{2\ell k} + \eta_{i\ell k} \tag{A.1}$$

$$u_{ij\ell k} = z_i\phi_{j\ell k} + \varepsilon_{ij\ell k} \tag{A.2}$$

$$V_{ij\ell k} = w_{i\ell k} + u_{ij\ell k} \tag{A.3}$$

$$w_{i\ell k} \text{ observed} \iff V_{ij\ell k} > V_{ij\ell' k'} \ \forall\,(\ell', k') \tag{A.4}$$

In the baseline model, I consider the case where ℓ comes from a 15-dimensional set, and where k is two-dimensional. Thus, there are 30 sectors in the model. For simplicity, s_i is a binary variable while x_i contains a mixture of binary and continuous variables. z_i contains a number of binary variables as well as two continuous exclusion restrictions which measure preference intensity for staying in the birth location and for working in the related occupation. In addition, z_i contains a number of interactions among this set of variables. $\eta_{i\ell k}$ is assumed to be distributed iid $N(0, 1)$ across all individuals, locations, and occupations. The same is true for $\varepsilon_{ij\ell k}$. In later simulations, I examine performance of the estimator when these error terms are correlated across locations and occupations.

The estimate of interest is $\hat{\gamma}_2$ in location 8 and occupation 2, which is chosen without loss of generality. The true value of this parameter is set to 2. I consider estimation of $\gamma_{2,8,2}$ in small samples (N=1,000 per sector) and large samples

(N=10,000 per sector). Each simulation is repeated 100 times, and I report the resulting bias and standard deviation of the parameter estimates, along with the average root mean squared error of the wage regression across repetitions.

I report the performance of nine different specifications under three different error structures. As a baseline, I include the naive OLS estimator that would be unbiased and consistent if no selection were present. I then consider four separate estimates of the selection probabilities in the polynomial selection terms. For each estimate, I consider including only the first-best probability, or the first-best and location probabilities as implemented in the empirical section of the paper. The four different probability estimators are as follows: (i) fully specified bin; (ii) conditional inference tree; (iii) logit; and (iv) coarse (misspecified) bin. I specifically include the coarse bin estimator to show the effect of the researcher being unable to include all relevant choice predictors, e.g. due to the curse of dimensionality. The three different error structures I consider are as follows: (i) the baseline described above; (ii) allowing the preference shocks to be correlated across locations and occupations (i.e. $\varepsilon_{ij\ell k}$ distributed iid $N(\mathbf{0}, \Sigma)$ across individuals, where Σ is a random covariance matrix); and (iii) allowing both preference shocks and earnings shocks to be multivariate normal distributions. Related work by Bourguignon, Fournier, and Gurgand (2007) assesses the properties of selection correction estimators when the preference shocks are iid logit.

The results of the simulations are reported in Table A1. Each of the three error structures are reported respectively in Panels A, B, and C of the table. Within each panel are the nine different specifications used to estimate $\gamma_{2,8,2}$. The main takeaway from the simulations is that the tree algorithm performs very similarly to the fully specified bin estimator in large samples, but that the tree algorithm performs much better than all other estimators in small samples. The improved small-sample performance of the tree algorithm is consistent with Asher et al. (2016), who prove the consistency of tree classification and also show excellent

small sample performance. For all specifications, the OLS estimate of the parameter of interest is severely downward biased, while the logit estimate is severely upward biased. The coarse bin estimator performs only slightly better than OLS and incurs a high efficiency cost.

The purpose of Panels A and B is to show that the nonparametric estimator used in this paper performs well when the distribution of preference shocks is either normal or multivariate normal. In both of these two scenarios, index sufficiency holds. In Panel C, however, index sufficiency is less likely to hold. In this case, none of the estimators is able to completely resolve the selection problem. However, the tree estimator performs best, again particularly in smaller samples.

Table A1: Monte Carlo simulation results

	10,000 Observations per Sector				1,000 Observations per Sector			
	Bias	Std. Dev.	Regression RMSE	Ave. Sample Size	Bias	Std. Dev.	Regression RMSE	Ave. Sample Size
Panel A: 30 sectors, baseline								
OLS	-0.3781	0.0308	0.8994	28502	-0.3819	0.0984	0.8974	2857
1st Best Bin	-0.0405	0.0304	0.8726		-0.0719	0.0967	0.8724	
1st Best Tree	-0.0406	0.0303	0.8704		0.0256	0.0998	0.8721	
1st Best Logit	0.1458	0.0336	0.8687		0.1401	0.1045	0.8675	
1st Best Coarse Bin	-0.1495	0.0344	0.8959		-0.1611	0.1071	0.8942	
1st+2nd Best Bin	-0.0587	0.0332	0.8723		-0.0915	0.1047	0.8720	
1st+2nd Best Tree	-0.0490	0.0329	0.8702		0.0125	0.1063	0.8716	
1st+2nd Best Logit	0.1523	0.0409	0.8684		0.1734	0.1248	0.8673	
1st+2nd Best Coarse Bin	-0.3814	0.1148	0.8944		-0.2557	0.2826	0.8931	
Panel B: 30 sectors, $\varepsilon_{ij\ell k}$ correlated across (ℓ, k)								
OLS	-0.3405	0.0311	0.9123	27504	-0.3061	0.0870	0.9136	2755
1st Best Bin	-0.0384	0.0340	0.8929		-0.0418	0.0874	0.8949	
1st Best Tree	-0.0408	0.0323	0.8916		0.0269	0.0896	0.8953	
1st Best Logit	0.1152	0.0361	0.8903		0.1406	0.0905	0.8912	
1st Best Coarse Bin	-0.1181	0.0372	0.9095		-0.1003	0.0967	0.9109	
1st+2nd Best Bin	-0.0549	0.0376	0.8927		-0.0526	0.1042	0.8947	
1st+2nd Best Tree	-0.0530	0.0338	0.8914		0.0102	0.1075	0.8949	
1st+2nd Best Logit	0.1131	0.0431	0.8901		0.1622	0.1325	0.8909	
1st+2nd Best Coarse Bin	-0.3263	0.1149	0.9085		-0.1940	0.2781	0.9100	
Panel C: 30 sectors, both $\varepsilon_{ij\ell k}$ and $\eta_{i\ell k}$ correlated across (ℓ, k)								
OLS	-0.4596	0.0943	1.0715	26613	-0.4638	0.1266	1.0679	2676
1st Best Bin	-0.0572	0.0513	1.0437		-0.1202	0.1031	1.0427	
1st Best Tree	-0.0616	0.0505	1.0419		-0.0326	0.1089	1.0424	
1st Best Logit	0.1334	0.0554	1.0400		0.1130	0.1145	1.0374	
1st Best Coarse Bin	-0.1297	0.0662	1.0670		-0.1687	0.1325	1.0639	
1st+2nd Best Bin	-0.0768	0.0534	1.0433		-0.1217	0.1218	1.0422	
1st+2nd Best Tree	-0.0757	0.0527	1.0416		-0.0496	0.1292	1.0418	
1st+2nd Best Logit	0.1329	0.0598	1.0397		0.1453	0.1685	1.0370	
1st+2nd Best Coarse Bin	-0.4195	0.1352	1.0658		-0.2881	0.3828	1.0628	

Note: 100 replications used for all specifications. "Regression RMSE" refers to the standard deviation of the regression residuals. This quantity is larger than unity in Panel C because of the correlation between $\varepsilon_{ij\ell k}$ and $\eta_{i\ell k}$. "OLS" indicates OLS estimation of the parameter of interest, ignoring potential selection bias. "1st Best Bin" indicates estimation of equation (3.5) using a cubic polynomial of the first-best probability from a simple bin estimator. "1st+2nd Best Bin" indicates the same, except that both the first-best and occupation probabilities are used, as described in Section 3.2. The polynomial is a full set of third-degree polynomial terms, including interactions. "Tree" refers to estimation using probabilities from the conditional inference tree algorithm described in Section 5.1.2. "Logit" indicates estimation using probabilities from a logit model. "Coarse Bin" refers to estimation using probabilities from a more coarsely defined bin estimator, as would be required in the empirical application of this paper.

B Online Appendix: Data Details

This section describes additional details relating to the construction of the earnings and demographic variables used in the analysis.

B.1 Overview of American Community Survey

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, demographic characteristics (e.g. age, gender, race, ethnicity, household composition), college major, advanced degree attainment, occupation, and earnings.

B.2 Aggregation of majors

The ACS records hundreds of distinct college major fields following the Classification of Instructional Programs (CIP) established by the National Center for Education Statistics (NCES). Altonji, Kahn, and Speer (2016) aggregate these into 51 majors. I then further aggregate these into groups with similar pre- and post-graduation outcome characteristics in order to focus the analysis and to maintain statistical power. A detailed mapping of the 51 Department of Education major fields to these five aggregated fields is provided in Table F2.

B.3 Definition of related occupations

Occupations are either *related* to the major or *unrelated*. As mentioned in the text, an occupation is 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).[9] 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. I allow the set of related occupations to differ based on advanced degree status.

The cutoff of 2% was chosen so as to ensure that highly specialized majors (i.e. majors with high concentration in few occupations) would have their most concentrated occupations defined as related. To provide further intuition for this approach, I present in Figure F1 the frequency distribution of occupations (sorted from most to least frequent) for non-advanced-degree holders in four majors: primary education, history, economics, and computer programming. For each panel of the figure, I include a vertical line along with the frequency distribution, which serves to mark the cutoff between related and unrelated occupations. Figure F1 shows that the primary education and computer programming majors are highly specialized, with 30%-40% of majors working in the most common occupation (elementary school teachers and software developers, respectively). Furthermore, computer programming majors are observed in many fewer occupations than the other majors included in the figure, by a factor of four. On the other hand, economics and history majors do not have clear-cut occupations corresponding to them, as the most frequent occupation contains only 10% of majors (miscellaneous managers for both). Figure F2 reports the same information but for advanced degree holders only. The results are similar. The exact occupation titles that are related to each of these majors are listed in Tables F3 and F4, respectively, by advanced degree status.

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

While the 2% cutoff for defining related occupations may seem arbitrary, the rule results in a construction of majors and occupations that aligns with common sense and other papers in the literature.[10] A list of related occupations for each of the five aggregate college major categories is included in Table F5 for bachelor's degree holders and Table F6 for advanced degree holders. 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 Tables F5 and F6. First, almost all engineering occupations are not considered to be related to any major except STEM.[11] Second, salespersons and 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 and authors tend to work in. Finally, accountants and auditors are related to business majors, other majors, and STEM majors. Of additional note is that Table F6 includes a set of occupations not included in Table F5 such as actuaries, pharmacists, and lawyers. These occupations all have the expected relatedness with bachelor's degree major: actuaries and pharmacists are related only to STEM, while lawyers are related to all majors except education.

[10] 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 (2015) 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.

Abel and Deitz (2015) pursue a different approach by mapping college majors to occupations using the Department of Labor's O*NET data and crosswalks provided by the Department of Education's National Center for Education Statistics (NCES). They distinguish between occupations that are a "college degree match" and occupations that are a "college major match" and find that college graduates with better job matches earn higher wages and that better matches are more likely to occur in larger labor markets. In a similar fashion, Freeman and Hirsch (2008) use O*NET data to map occupational skills to majors.

[11] Civil engineers and industrial engineers are also related to the "other" category of majors.

Based on this set of illustrative examples, the definition of occupation relatedness posed here is reasonable.

B.4 Definitions and details on ACS variables used

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.

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."

C Online Appendix: Model Simplification

The model described in the paper makes a number of simplifying assumptions, which I discuss here at greater length. I also discuss other potential data sources that could be used for an analysis of migration and college major choice. Finally, I present estimates on the rate of out-of-state college attendance and out-of-state migration after college.

C.1 Simplifying assumptions

The two main simplifying assumptions are: (i) not treating college major as a decision; and (ii) restricting migration and occupational choice to be "once-and-for-all."

These simplifying assumptions are driven primarily by data restrictions in the American Community Survey (ACS). Without details about a student's abilities, past location choices, what type of college he attended, or where that college was located, it is impossible to credibly enrich the model to account for selective major choice or dynamics in migration and occupational choices.

C.1.1 No selection into major

Without data on cognitive and non-cognitive abilities, family background, or university identity, it is impossible to model selection into majors. Thus, the model treats majors as exogenous.

This modeling assumption likely results in overstated estimates of the returns to major. Indeed, my usage of the phrase "corrected return" is somewhat of a misnomer in the absence of modeling major choice. However, I retain this usage for comparability with Dahl (2002), who also adopts such usage.

It is unclear the extent to which abstracting from major choice affects the "uncorrected" (for location and occupational choices) and "corrected" returns. This

is because both migration and occupational choices are correlated with the same factors that determine major choice: cognitive and non-cognitive abilities, family background, university characteristics, etc. That said, it is likely that the "corrected" returns would be even lower if selectivity in major choice were appropriately corrected for.

C.1.2 Location and occupation choices are once-and-for-all

I employ an even stronger assumption in my model that location and occupational choices are made once-and-for-all at the conclusion of college.

This assumption is primarily motivated by lack of data on where the student graduated high school, and where he attended college, in addition to where he was born and where he lives as an adult. A more appropriate model of the selection process would allow for multiple move decisions. At a minimum, if the location of the university were known, I could estimate transition probabilities between birth state, university state, and adult residence state. For example, I could incorporate a variable indicating "attended college in this state" or "attended high school in this state" into the estimation of the selection probabilities. With more frequent (e.g. annual) panel data, a more formal approach could adopt a dynamic model like Kennan and Walker (2011), although that approach puts more focus on the parameters of the selection equation as opposed to the outcome equation. As an example, work by Kennan (2020) examines the interaction between migration and college completion in a dynamic setting using the National Longitudinal Survey of Youth 1979 (NLSY79), but is unable to capture heterogeneity across majors because of data limitations.

What are the pros and cons of abstracting from out-of-state college attendance? The primary gain is computational tractability. The primary loss is misspecification of the selection process. If graduates of certain majors are more likely to have attended college out-of-state, then excluding this information from the selection

equation can result in biased estimates of the "corrected" returns.

C.1.3 Other assumptions

The model also assumes that individuals have no uncertainty regarding their earnings or tastes in other locations. While it is possible to allow for imperfect information, doing so would require, e.g. assuming that the individual's information set is shared by the econometrician. On the other hand, the approach taken here to model migration in response to individual earnings shocks departs from much of the migration literature, which assumes that migration decisions are influenced by the deterministic portion of earnings (Kennan and Walker, 2011; Bishop, 2012; Ransom, Forthcoming). This assumption is typically made for tractability of dynamic models.

C.2 Other data sources

Besides the ACS, there are three other data sources that track individuals' location history and college major: the National Longitudinal Survey of Youth (NLSY), the Panel Study of Income Dynamics (PSID), and the Survey of Income and Program Participation (SIPP). Additionally, the Integrated Postsecondary Education Data System (IPEDS) tracks how many students attend college in their state of residence.[12]

The drawback of these other data sources for the modeling framework of the current paper is sample size and/or panel length. The NLSY and PSID are long panels, but have small cross-sectional sample sizes. The SIPP has a larger cross-section, but its panel length is only four years. IPEDS is a panel data set where the sampling unit is a university.

Aside from sample size concerns, detailed location information (such as state of

[12]The US Census Bureau has an experimental product called the Post-Secondary Employment Outcomes (PSEO), but this product does not (yet) have migration flows by college major.

residence) is typically redacted from publicly accessible extracts of the NLSY, PSID and SIPP. Obtaining data access for these sensitive variables requires an additional application process.

C.3 Estimates of out-of-state college attendance and college major choice

IPEDS reports the percentage of students enrolled in college in each state who are residents of that state.[13] In the average US state, 72% of enrolled students are residents of that state. However, this masks substantial heterogeneity: that number is over 90% for three states (Alaska, Texas and New Jersey) and less than 45% for three states (Rhode Island, Vermont and New Hampshire). Less than 31% of New Hampshire college students are residents of New Hampshire.

To supplement the evidence on out-of-state attendance, I collected data from the University of Oklahoma (OU) on graduating majors and in- vs. out-of-state status. OU is the flagship public institution for the state of Oklahoma. About 40% of its students come from out of state, with the vast majority from Texas. Table C1 shows the number of a cohort of enrolled students who graduated in each major, by residency status. Aside from Business, there does not appear to be any differential selection into broad majors by residency status.

C.4 Estimates of out-of-state migration after college

Data on post-college migration is extremely scarce. I am aware of only three sources that can produce estimates of post-college out-of-state migration rates: the NLSY, the PSID, and websites economicmodeling.com (EMSI) and democratizeopportunity.com. The two survey sources collect information on college major and repeatedly collect location information over an extended period of time. The two website sources rely

[13] See https://nces.ed.gov/ipeds/Search?query=residence&query2=residence&resultType=all&page=1&sortBy=date_desc&overlayTableId=26395.

Table C1: Out-of-state enrollment by completed major, University of Oklahoma

Major	N out-of-state	N in-state	% out-of-state
Education	256	374	40.63
Social Sciences	1,401	2,621	34.83
Other	2,365	3,412	40.94
Business	2,152	2,033	51.42
STEM	2,828	4,831	36.92
Total	9,002	13,271	40.42

Note: This table shows out-of-state enrollment by major for one cohort at the University of Oklahoma.

Source: University of Oklahoma, via Brent Norwood.

on data from resumes or job applications to infer location of college attended and current location. They do not report differences by college major.

EMSI identifies Texas, California, Washington, Georgia and Florida as the top five states in terms of retaining their college graduates. It finds New Hampshire, Vermont, Rhode Island, West Virginia and Wyoming to be the bottom five states in terms of retaining college graduates.

The top five states according to EMSI all fall in the top quartile of in-state college attendance according to IPEDS, while the bottom five states all fall in the bottom quartile of in-state college attendance. Thus, there is a strong correlation between how attractive a state's higher education system is, and how attractive that state is for post-college life.

D Online Appendix: Testing for Equality in Uncorrected and Corrected Returns

One important aspect of my analysis is determining whether the uncorrected and corrected coefficients are significantly different from one another. In this appendix, I describe how to assess whether my estimates of the corrected returns "matter" in the sense of being statistically and economically significantly different.

D.1 Intuition from a textbook example

I first provide intuition using an illustrative textbook example of omitted variable bias in a setting with one observed regressor x_1 and one unobserved regressor x_2. Assume that including x_2 would resolve the problem of omitted variable bias. In my model, x_1 would be a dummy for a particular major, and x_2 would represent the correction function (a polynomial of the selection probabilities). y represents the log earnings.

Now consider the fitted values from uncorrected and corrected models:

$$\widetilde{y} = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_1 \tag{D.1}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{D.2}$$

where the tilde simply emphasizes that the actual estimates are potentially different across the two. In my model, $\widetilde{\beta}_1$ represents the uncorrected wage return to a given major, while $\hat{\beta}_1$ represents the corrected return.

A natural question would then be to ask under what conditions the $\widetilde{\beta}_1$ would have a different estimate from $\hat{\beta}_1$, i.e. under what conditions are the estimates in (D.1) unbiased? The answer depends on whether x_2 and x_1 covary, and on whether

x_2 and y covary. In the simple example where x_1 and x_2 are scalars, we have

$$\widetilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \widetilde{\delta}_1 \tag{D.3}$$

where $\widetilde{\delta}_1$ is the estimated coefficient from a regression of x_2 on x_1. Thus, in order for $\widetilde{\beta}_1 \neq \hat{\beta}_1$, both $\hat{\beta}_2 \neq 0$ and $\widetilde{\delta}_1 \neq 0$.

D.2 Using F and Hausman tests in my model

In a more general model like the one I present, (D.3) does not hold, since x_1 and x_2 are both vector-valued. In the more general case, the condition that $\hat{\beta}_2 \neq 0$ generalizes to a condition that the F statistic for joint significance of the correction probabilities is larger than the critical value.[14] The condition that $\widetilde{\delta}_1 \neq 0$ is more difficult to directly analogize, but it captures the situation where the polynomial of correction probabilities is correlated with the major dummies.

A more straightforward way of assessing whether the corrected and uncorrected returns are different is to directly compare their difference. The most natural way of doing this is a Hausman (1978) test, as implemented by Dahl (2002). Under the null hypothesis of no selection bias, the uncorrected returns estimated by OLS are consistent and efficient, and the corrected returns are consistent and inefficient. Under the alternative hypothesis, OLS is inconsistent but the corrected returns are consistent. Under these conditions, the variance of the difference in the estimates is the difference in the variances.

[14] I compute the F statistic by bootstrap. For each bootstrap replicate, I compute the formula

$$F = \frac{(SSR_{\text{uncorr}} - SSR_{\text{corr}})/q}{SSR_{\text{corr}}/(N - K_{\text{corr}})} \tag{D.4}$$

and then average across all 500 bootstrap replicates, where SSR is the sum of the squared residuals, q denotes the number of correction terms (equal to 6), and K_{corr} is the number of regressors in the corrected model.

To obtain the associated p-value, I average the frequency with which the F statistic exceeds the critical value for a $F(q, N - K_{\text{corr}})$ distribution at the 5% level of significance.

The Hausman test statistic is given by

$$\begin{aligned} H &= \left(\beta_{\text{uncorr}} - \beta_{\text{corr}}\right)' \left(\text{Var}(\beta_{\text{corr}}) - \text{Var}(\beta_{\text{uncorr}})\right)^{-1} \left(\beta_{\text{uncorr}} - \beta_{\text{corr}}\right) \\ &\sim \chi^2_K \end{aligned} \tag{D.5}$$

where K denotes the degrees of freedom of the χ^2 distribution, which in this case is the rank of the differenced β vector. The variance terms are computed using 500 bootstrap replications, as discussed in the paper.

I compute the Hausman test statistic for each major interacted with advanced degree status. The uncorrected return, corrected return, percentage difference after correcting for selection, Hausman statistic and p-value, and F statistic and p-value are listed in Tables F10–F27 for each location and each major and advanced degree.

D.3 Assessing statistical and economic significance

While the approach described above may result in statistically different returns, the difference in the returns may not be economically significant. For example, in Table 2, the estimated returns to working in an unrelated occupation as an education major with an advanced degree are 0.034 (uncorrected) and 0.023 (corrected). Table F18 indicates that the Hausman statistic for the difference in these is 10.7 with a corresponding p-value of 0.001. The question is then whether a reduction of 1.1 log points should be considered economically significant, especially when neither the uncorrected nor corrected estimates are not statistically different from zero.

I report in Table 3 moments of the distribution of percentage differences between corrected and uncorrected returns. Of particular importance is the column labeled "No. significant," which counts the number of locations that I label as being "statistically and economically significant." To be labeled such, the estimate in question must satisfy the following conditions: (i) the Hausman test statistic overturns the null hypothesis of no difference between corrected and uncorrected

at the 5% level; (ii) both the uncorrected and corrected coefficients are statistically different from zero at the 5% level; (iii) the percentage difference between the corrected and uncorrected returns is significantly different from zero at the 5% level; and (iv) the percentage difference exceeds 10% in magnitude.

To more readily visualize these differences, I also produce Figures F3–F6 which plot the uncorrected and corrected returns for each major and advanced degree level, and for each location. Points that are statistically and economically significant are colored black.

Selection matters the most for advanced degrees, and particularly for those in a related occupation. This makes sense given that the individuals with these skills have invested more resources to become more specialized. Another interesting finding is that selection seems to matter much more for locations in the Northeast (e.g. New England, New York, and New Jersey/Pennsylvania) than in other parts of the US.

E Online Appendix: Further Details on Tree Estimation

This appendix provides more detail on the estimation and results of the conditional inference tree algorithm.

E.1 Estimates of the choice probabilities

The algorithm partitions the set of covariates in order to maximize predictive classification accuracy of the dependent variable. It 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:[15]

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

E.2 Estimates of the choice probabilities

In discretizing the probabilities, I assume that all individuals in a given cell have the same choice probability.[16] Thus, deviations from the cell mean correspond to a reduced-form measure of preference shocks, which allow me to separate preferences from earnings. Because of their key role in identification, I present in Table F9 moments of the distributions of average cell probabilities, conditional on major, occupation, and move-stay decision. The table also reports the number of individuals in each migration-occupation-major classification and the number of different cells contributing to each classification.

[15]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 1% for the p-value parameter, 500 observations for the minimum cluster size, and 500 observations for the minimum candidate node split size. I also adjust the p-value for multiple comparisons. I also tried other permutations of these parameters and found that the overall performance of the tree algorithm was not sensitive to these tuning parameters.

[16]For parametric choice models, the analogous assumption is that the choice probability is the same for all individuals with the same values for all covariates.

In results available upon request, the preference shifters discussed in the paper have a strong first-stage. These shifters appear relatively high up in the tree and also appear in many different splits. The tree algorithm described above allows for these shifters to enter highly non-linearly into the first stage.

The probabilities listed in Table F9 also confirm the earlier descriptive analysis of Figures 1 and 2. The cell probabilities in panels (a) and (c), which correspond to working in a related occupation, are highest among education, business, and STEM majors. Another way to see this is to compare the difference in average cell probabilities for working in a related occupation relative to working in an unrelated occupation. This difference is much higher for education, business, and STEM majors than for the remaining two.

Finally, note that the number of cells is larger for movers than for stayers, and that the number of cells roughly corresponds to the number of individuals within a major-occupation category. The difference in the number of cells is much less stark than if a bin estimator were to be used, because the tree algorithm automatically merges together sparse bins, or bins that are not statistically distinct, to avoid overfitting.

Online Appendix References

Abel, Jaison R. and Richard Deitz. 2015. "Agglomeration and Job Matching among College Graduates." *Regional Science and Urban Economics* 51:14–24.

Altonji, Joseph G., Lisa B. Kahn, and Jamin D. Speer. 2016. "Cashier or Consultant? Entry Labor Market Conditions, Field of Study, and Career Success." *Journal of Labor Economics* 34 (S1):S361–S401.

Asher, Sam, Denis Nekipelov, Paul Novosad, and Stephen P. Ryan. 2016. "Classification Trees for Heterogeneous Moment-Based Models." Working Paper 22976, National Bureau of Economic Research.

Bishop, Kelly. 2012. "A Dynamic Model of Location Choice and Hedonic Valuation." Working paper, Washington University in St. Louis.

Bourguignon, François, Martin Fournier, and Marc Gurgand. 2007. "Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons." *Journal of Economic Surveys* 21 (1):174–205.

Dahl, Gordon B. 2002. "Mobility and the Return to Education: Testing a Roy Model with Multiple Markets." *Econometrica* 70 (6):2367–2420.

Freeman, James A. and Barry T. Hirsch. 2008. "College Majors and the Knowledge Content of Jobs." *Economics of Education Review* 27 (5):517–535.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 67 (6):1251–1272.

Kennan, John. 2020. "Spatial Variation in Higher Education Financing and the Supply of College Graduates." Working paper, University of Wisconsin-Madison.

Kennan, John and James R. Walker. 2011. "The Effect of Expected Income on Individual Migration Decisions." *Econometrica* 79 (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 (4):933–972.

Ransom, Tyler. Forthcoming. "Labor Market Frictions and Moving Costs of the Employed and Unemployed." *Journal of Human Resources* .

F Online Appendix: Supporting Figures and Tables

Table F1: Sample selection details

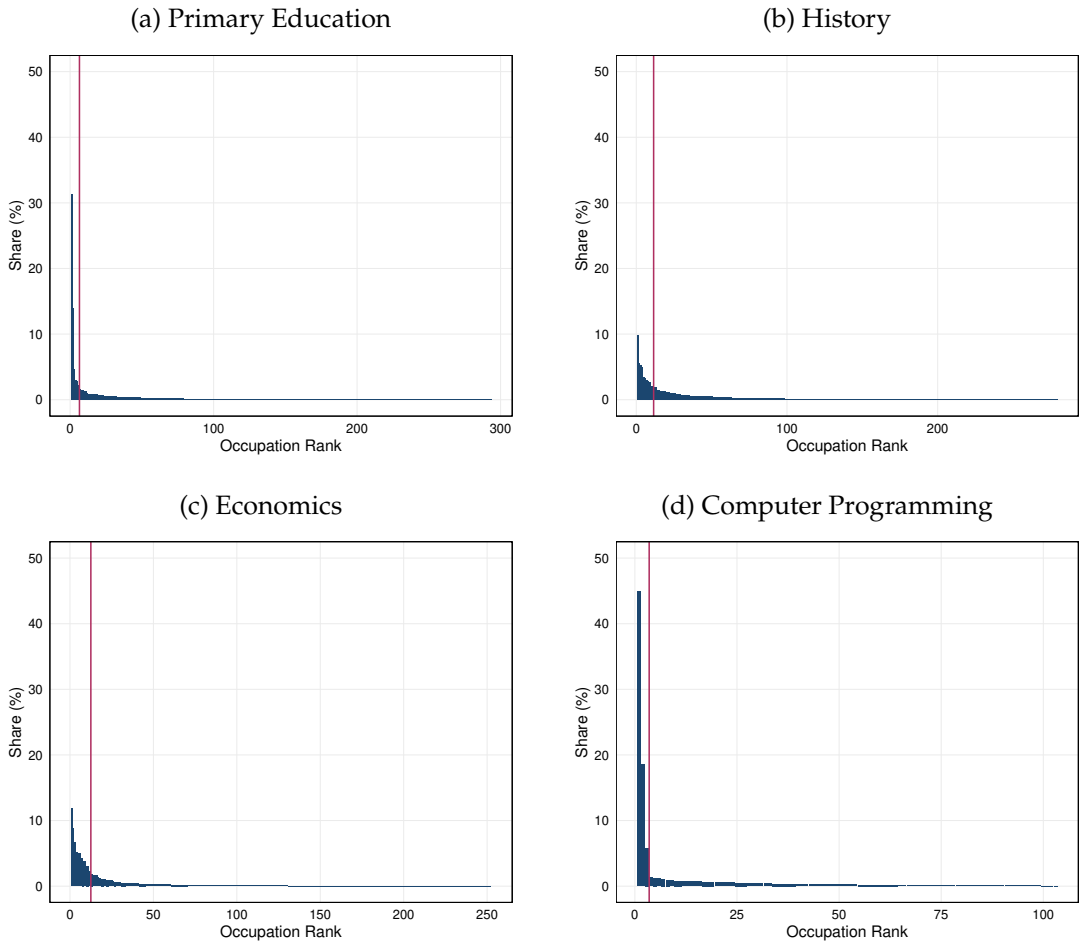
Criterion	No. obs deleted	Remaining obs.
Respondents in 2010-2019 ACS	—	31,499,768
Drop those without a bachelor's degree or higher	24,355,024	7,144,744
Drop those outside of 22-54 age range	2,828,575	4,316,169
Drop those currently enrolled in school or living in group quarters	479,430	3,836,739
Drop those not born in the US	695,962	3,140,777
Drop those with missing annual earnings	0	3,140,777
Drop those with positive annual earnings below $20,000	318,123	2,822,654
Drop those with annual earnings above $600,000	4,217	2,818,437
Drop those with zero annual earnings	353,406	2,465,031
Drop those with missing occupation	0	2,465,031
Drop females	1,258,967	1,206,064
Drop those with imputed earnings or occupations	180,561	1,025,503
Drop those with imputed labor force status	1,978	1,023,525
Final analysis sample	—	1,023,525

Table F2: 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	

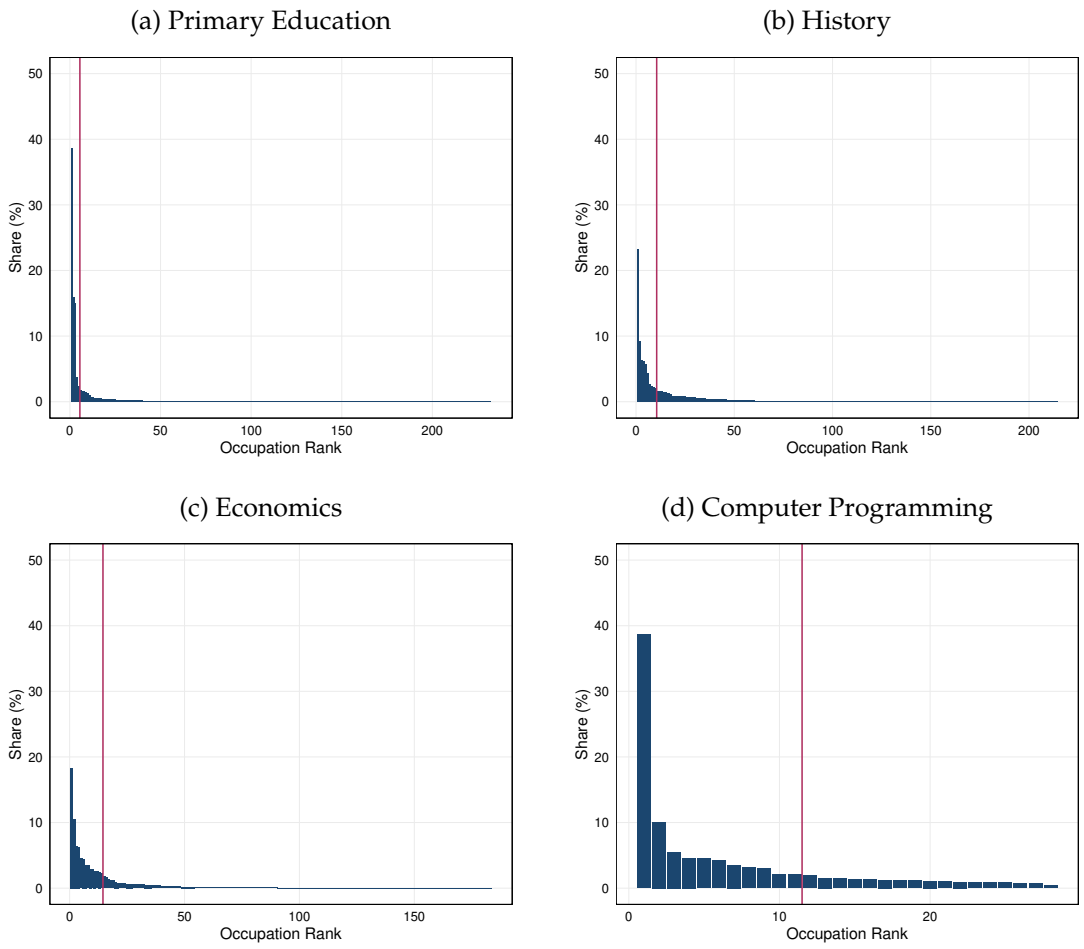
Note: Aggregation of the 51 detailed Department of Education majors analyzed in Altonji, Kahn, and Speer (2016).

Figure F1: Occupation distributions for select detailed majors: Non-advanced degree holders



Notes: Graphs represent occupation distributions conditional on detailed major. Vertical lines represent the cutoff between related and unrelated occupations: those to the left of the line are related, while those to the right are unrelated. Additional details regarding the definition of occupation relatedness are provided in the text and the appendix.
Source: Author's calculations from American Community Survey, 2010-2019.

Figure F2: Occupation distributions for select detailed majors: Advanced degree holders



Notes: Graphs represent occupation distributions conditional on detailed major. Vertical lines represent the cutoff between related and unrelated occupations: those to the left of the line are related, while those to the right are unrelated. Additional details regarding the definition of occupation relatedness are provided in the text and the appendix.
Source: Author's calculations from American Community Survey, 2010-2019.

Table F3: List of frequent occupations for select majors: Non-advanced degree holders

(a) Primary Education

Occupation	Share (%)
Primary school teachers	31.37
Secondary school teachers	13.92
Managers and administrators, n.e.c.	4.63
Supervisors and proprietors of sales jobs	2.92
Salespersons, n.e.c.	2.84
Teachers , n.e.c.	2.16
Police, detectives, and private investigators	1.62
Computer systems analysts and computer scientists	1.46
Retail sales clerks	1.39

(b) History

Occupation	Share (%)
Managers and administrators, n.e.c.	9.80
Salespersons, n.e.c.	5.54
Supervisors and proprietors of sales jobs	5.30
Primary school teachers	4.92
Computer systems analysts and computer scientists	3.47
Military	3.25
Police, detectives, and private investigators	2.97
Secondary school teachers	2.76
Managers and specialists in marketing, advertising, and public relations	2.71
Customer service reps, investigators and adjusters, except insurance	2.07
Other financial specialists	2.06
Retail sales clerks	1.97
Chief executives and public administrators	1.89
Financial managers	1.47

(c) Economics

Occupation	Share (%)
Managers and administrators, n.e.c.	11.79
Other financial specialists	8.80
Salespersons, n.e.c.	6.63
Supervisors and proprietors of sales jobs	5.09
Financial managers	5.03
Computer systems analysts and computer scientists	4.96
Accountants and auditors	4.23
Financial services sales occupations	3.86
Chief executives and public administrators	3.80
Managers and specialists in marketing, advertising, and public relations	3.07
Management analysts	3.05
Computer software developers	2.21
Customer service reps, investigators and adjusters, except insurance	1.86
Retail sales clerks	1.60
Insurance sales occupations	1.60

(d) Computer Programming

Occupation	Share (%)
Computer software developers	44.86
Computer systems analysts and computer scientists	18.50
Managers and administrators, n.e.c.	5.82
Accountants and auditors	1.30
Repairers of data processing equipment	1.25
Managers and specialists in marketing, advertising, and public relations	1.15

Notes: Tables list occupations within the given major that are above the 2% cutoff defining relatedness, along with three additional occupations below the cutoff.

Table F4: List of frequent occupations for select majors: Advanced degree holders

(a) Primary Education

Occupation	Share (%)
Primary school teachers	38.70
Managers in education and related fields	15.98
Secondary school teachers	14.97
Subject instructors (HS/college)	3.67
Managers and administrators, n.e.c.	2.29
Special education teachers	1.77
Teachers , n.e.c.	1.62
Vocational and educational counselors	1.61

(b) History

Occupation	Share (%)
Lawyers	23.28
Primary school teachers	9.26
Subject instructors (HS/college)	6.27
Managers and administrators, n.e.c.	6.23
Secondary school teachers	5.68
Managers in education and related fields	4.31
Military	2.62
Physicians	2.33
Clergy and religious workers	2.18
Chief executives and public administrators	2.09
Computer systems analysts and computer scientists	1.63
Managers and specialists in marketing, advertising, and public relations	1.57
Other financial specialists	1.56

(c) Economics

Occupation	Share (%)
Lawyers	18.34
Managers and administrators, n.e.c.	10.52
Financial managers	6.44
Other financial specialists	6.28
Chief executives and public administrators	4.52
Accountants and auditors	4.47
Management analysts	3.55
Subject instructors (HS/college)	3.50
Computer systems analysts and computer scientists	2.92
Supervisors and proprietors of sales jobs	2.80
Economists, market researchers, and survey researchers	2.60
Physicians	2.57
Managers and specialists in marketing, advertising, and public relations	2.46
Salespersons, n.e.c.	2.22
Financial services sales occupations	1.82
Primary school teachers	1.59
Managers in education and related fields	1.35

(d) Computer Programming

Occupation	Share (%)
Computer software developers	38.68
Computer systems analysts and computer scientists	10.09
Primary school teachers	5.52
Designers	4.62
Wood lathe, routing, and planing machine operators	4.53
Managers and administrators, n.e.c.	4.21
Supervisors and proprietors of sales jobs	3.52
Chief executives and public administrators	3.15
Industrial engineers	3.00
Subject instructors (HS/college)	2.06
Auto body repairers	2.05
Retail sales clerks	1.90
Lawyers	1.53
Police, detectives, and private investigators	1.48

Notes: Tables list occupations within the given major that are above the 2% cutoff defining relatedness, along with three additional occupations below the cutoff.

Table F5: Complete list of related occupations by major: Non-advanced degree holders

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 of medicine and health occupations					✓
Managers of food-serving and lodging establishments			✓		
Funeral directors			✓		
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.					✓
Agricultural and food scientists					✓
Biological scientists					✓
Foresters and conservation scientists					✓
Registered nurses					✓
Pharmacists					✓
Respiratory therapists					✓
Occupational therapists					✓
Physical therapists					✓
Therapists, n.e.c.					✓
Primary school teachers	✓	✓	✓		✓
Secondary school 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					✓
Radiologic tech specialists					✓
Health technologists and technicians, n.e.c.			✓		✓
Engineering technicians, n.e.c.					✓
Drafters			✓		
Chemical technicians					✓
Airplane pilots and navigators			✓		
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		✓	✓	✓	✓
Customer service reps, investigators and adjusters, except insurance		✓	✓	✓	
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		✓	✓		
Farmers (owners and tenants)					✓
Farm workers					✓
Supervisors of agricultural occupations					✓
Gardeners and groundskeepers					✓
Production supervisors or foremen					✓
Military		✓	✓		✓

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

Table F6: Complete list of related occupations by major: Advanced degree holders

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		✓	✓		
Architects			✓		
Aerospace engineer					✓
Chemical engineers					✓
Civil engineers					✓
Electrical engineer					✓
Industrial engineers					✓
Mechanical engineers					✓
Not-elsewhere-classified engineers					✓
Computer systems analysts and computer scientists		✓	✓	✓	✓
Operations and systems researchers and analysts		✓			
Actuaries					✓
Mathematicians and mathematical scientists					✓
Physicists and astronomers					✓
Chemists					✓
Atmospheric and space scientists					✓
Geologists					✓
Physical scientists, n.e.c.					✓
Agricultural and food scientists					✓
Biological scientists					✓
Foresters and conservation scientists					✓
Medical scientists			✓		✓
Physicians		✓	✓	✓	✓
Dentists		✓			✓
Veterinarians					✓
Other health and therapy					✓
Registered nurses					✓
Pharmacists					✓
Physical therapists					✓
Speech therapists					✓
Therapists, n.e.c.		✓			
Physicians assistants					✓
Subject instructors (HS/college)	✓	✓	✓	✓	✓
Primary school teachers	✓	✓	✓	✓	✓
Secondary school teachers	✓	✓	✓		✓
Teachers , n.e.c.			✓		✓
Vocational and educational counselors		✓	✓		
Archivists and curators			✓		
Economists, market researchers, and survey researchers				✓	
Psychologists		✓			
Urban and regional planners			✓		
Social workers		✓	✓		
Clergy and religious workers		✓	✓		
Lawyers		✓	✓	✓	✓
Designers			✓		✓
Musician or composer			✓		
Actors, directors, producers			✓		
Art makers: painters, sculptors, craft-artists, and print-makers			✓		
Editors and reporters			✓		
Athletes, sports instructors, and officials					✓
Clinical laboratory technologies and technicians					✓
Radiologic tech specialists					✓
Health technologists and technicians, n.e.c.			✓		✓
Airplane pilots and navigators			✓		
Computer software developers			✓		✓
Supervisors and proprietors of sales jobs		✓	✓	✓	✓
Financial services sales occupations				✓	
Salespersons, n.e.c.		✓	✓	✓	✓
Fire fighting, prevention, and inspection			✓		
Police, detectives, and private investigators		✓	✓		
Guards, watchmen, doorkeepers			✓		
Cooks, variously defined			✓		
Welfare service aides			✓		
Farmers (owners and tenants)					✓
Auto body repairers					✓
Production supervisors or foremen					✓
Wood lathe, routing, and planing machine operators					✓
Military		✓	✓	✓	✓

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

Table F7: Aggregation of locations

Location	2010 Population
California	39,144,818
OH, IN, MI, WI	33,927,016
Texas	27,469,114
NC, SC, GA	25,153,808
Mountain Census Division	23,530,498
NJ, PA	21,760,516
West North Central Census Division	21,120,392
Florida	20,271,272
New York	19,795,791
East South Central Census Division	18,876,703
WV, VA, DC, MD, DE	17,851,684
New England Census Division	14,727,584
AK, HI, OR, WA	13,369,363
Illinois	12,859,995
OK, AR, LA	11,560,266

Notes: The Mountain Census Division includes the following states: AZ, NM, CO, UT, NV, ID, MT, WY. The West North Central Census Division includes the following states: ND, SD, NE, KS, MO, IA, and MN. The East South Central Census Division is comprised of AL, MS, TN, and KY. The New England Census Division is comprised of CT, RI, MA, VT, NH, and ME.

Table F8: Predictive performance of various algorithms

	Classification algorithm		
Performance Criterion	Logit	Bin	Tree
Training set performance:			
Accuracy	37.45%	36.01%	38.19%
Kappa	34.67%	33.19%	35.45%
Test set performance:			
Accuracy	37.21%	35.37%	37.60%
Kappa	34.42%	32.52%	34.83%

Note: "Logit" refers to a flexibly specified logit; "Bin" refers to a 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 2010-2019 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.
Accuracy $= \frac{\text{number of correctly classified predictions}}{\text{number of predictions}}$.
Kappa $= \frac{\text{Accuracy} - \text{Expected Accuracy}}{1 - \text{Expected Accuracy}}$.
Expected Accuracy is defined as 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 (for example, such strategies could yield high accuracy for classification problems where one class is extremely rare).

Table F9: Summary of cell probabilities of observed decisions

(a) Stayers, Related occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	277	24,321	0.5138	0.1531	0.2824	0.6970
Social Sciences Major	270	28,348	0.3483	0.1141	0.2008	0.4897
Other Major	270	57,967	0.3505	0.1044	0.2082	0.4704
Business Major	310	96,094	0.4258	0.1093	0.2601	0.5392
STEM Major	323	109,439	0.4025	0.1104	0.2356	0.5174

(b) Stayers, Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	392	10,238	0.2847	0.1227	0.1301	0.4385
Social Sciences Major	381	28,979	0.3239	0.1263	0.1523	0.4813
Other Major	391	50,740	0.3123	0.1055	0.1619	0.4392
Business Major	410	58,575	0.2730	0.0890	0.1538	0.3786
STEM Major	407	65,172	0.2573	0.0895	0.1374	0.3644

(c) Movers, Related occupation

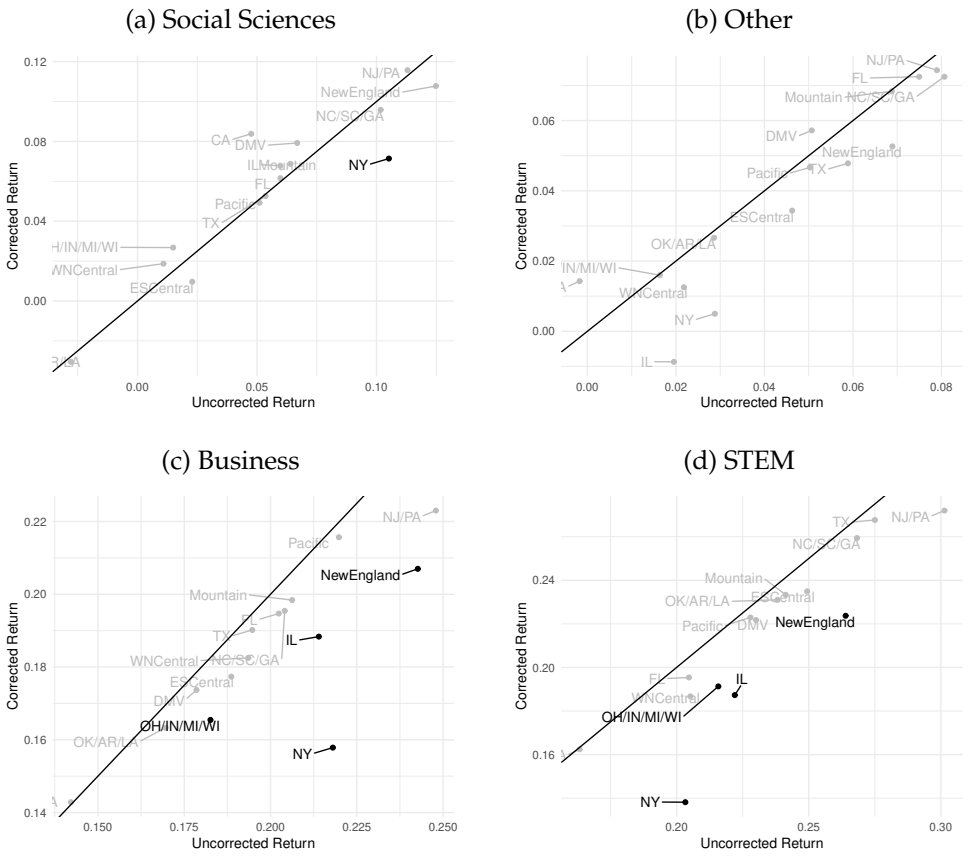
Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	673	12,305	0.1720	0.1899	0.0083	0.4576
Social Sciences Major	759	30,988	0.1016	0.1215	0.0089	0.2760
Other Major	766	58,867	0.0999	0.1202	0.0092	0.2690
Business Major	814	80,055	0.1296	0.1514	0.0088	0.3535
STEM Major	851	117,588	0.1184	0.1403	0.0103	0.3180

(d) Movers, Unrelated occupation

Education Level	Cells	Individuals	Mean	Std. Dev.	10th Percentile	90th Percentile
Education Major	666	7,307	0.0920	0.1137	0.0047	0.2498
Social Sciences Major	720	29,278	0.0842	0.1052	0.0067	0.2325
Other Major	753	48,298	0.0814	0.1024	0.0066	0.2249
Business Major	762	45,247	0.0765	0.0960	0.0052	0.2169
STEM Major	786	63,719	0.0702	0.0877	0.0057	0.2026

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, 2010-2019.

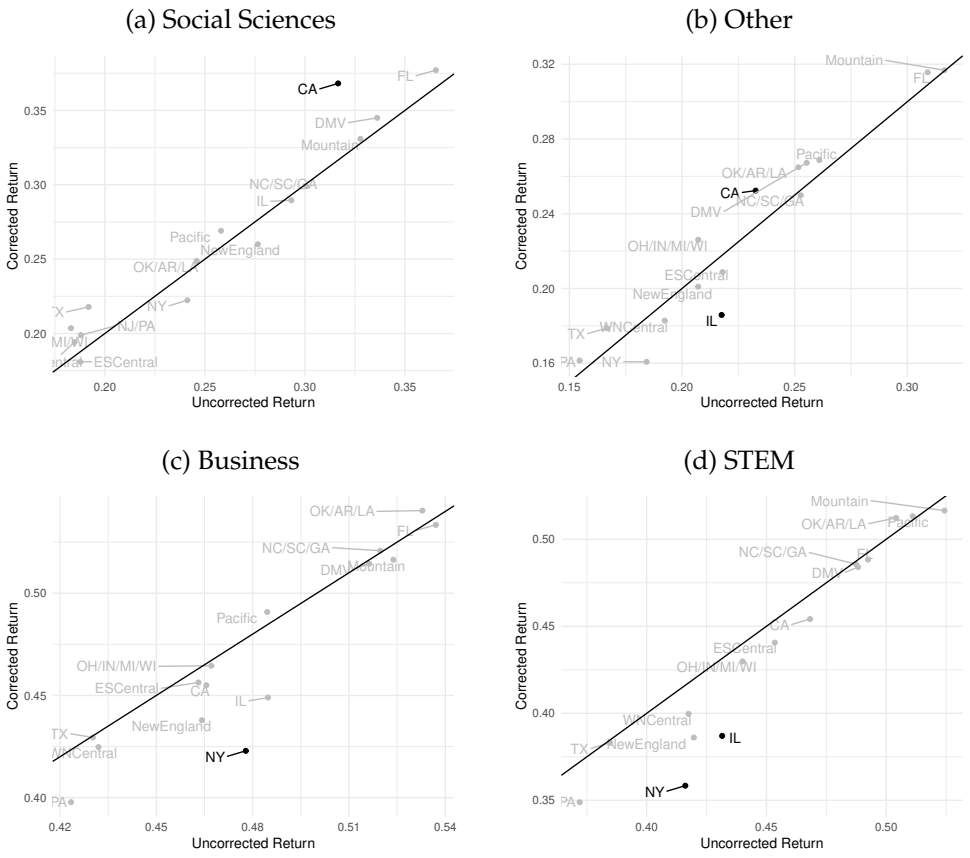
Figure F3: Scatter plots of uncorrected and corrected returns to major and working in an unrelated occupation



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Dots mark the uncorrected and corrected returns. Gray colored dots and labels indicate statistically and or economically insignificant differences at the 5% level. See Appendix D for further details.

Source: Author's calculations from American Community Survey, 2010-2019.

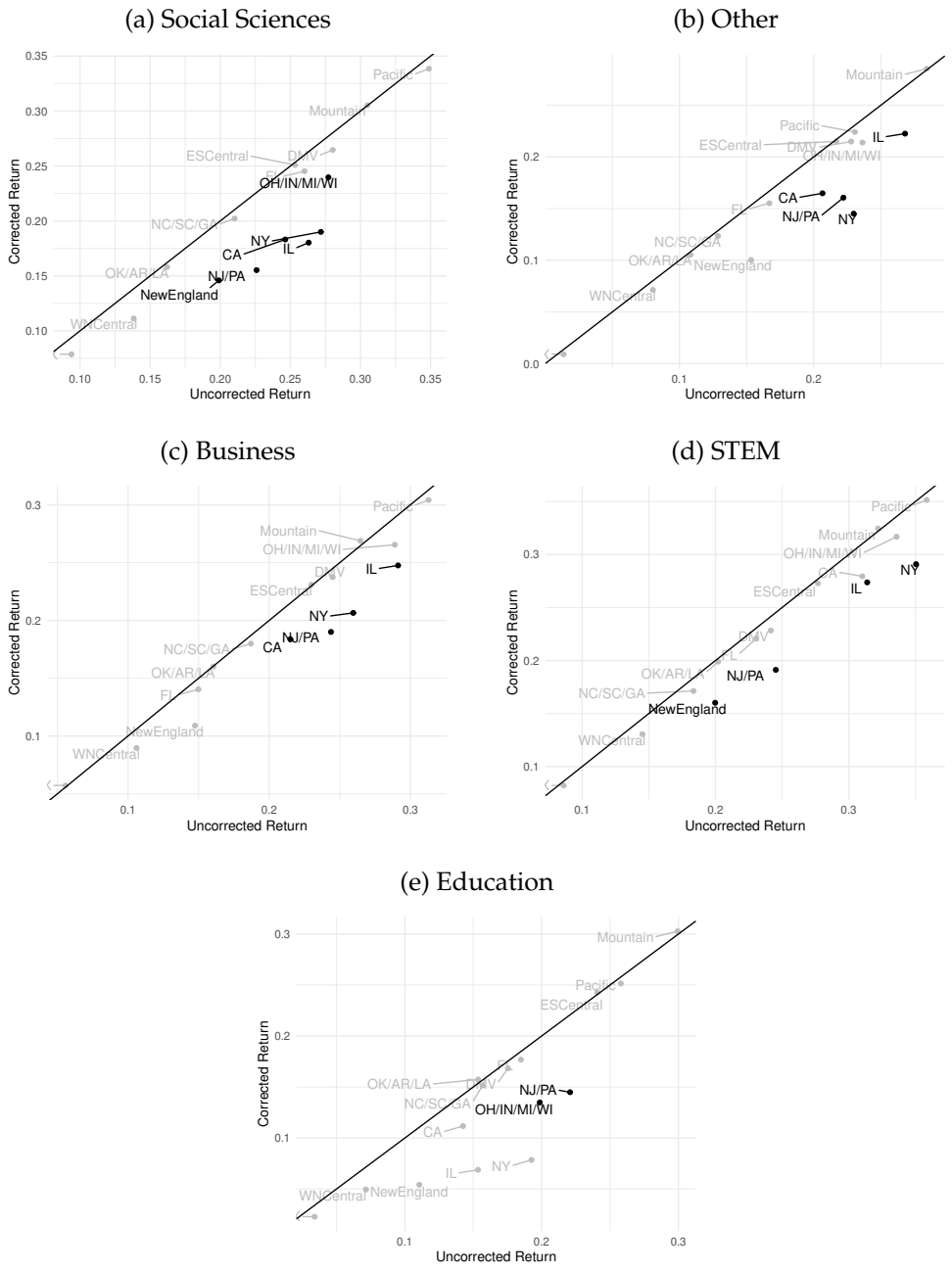
Figure F4: Scatter plots of uncorrected and corrected returns to major and working in a related occupation



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Dots mark the uncorrected and corrected returns. Gray colored dots and labels indicate statistically and or economically insignificant differences at the 5% level. See Appendix D for further details.

Source: Author's calculations from American Community Survey, 2010-2019.

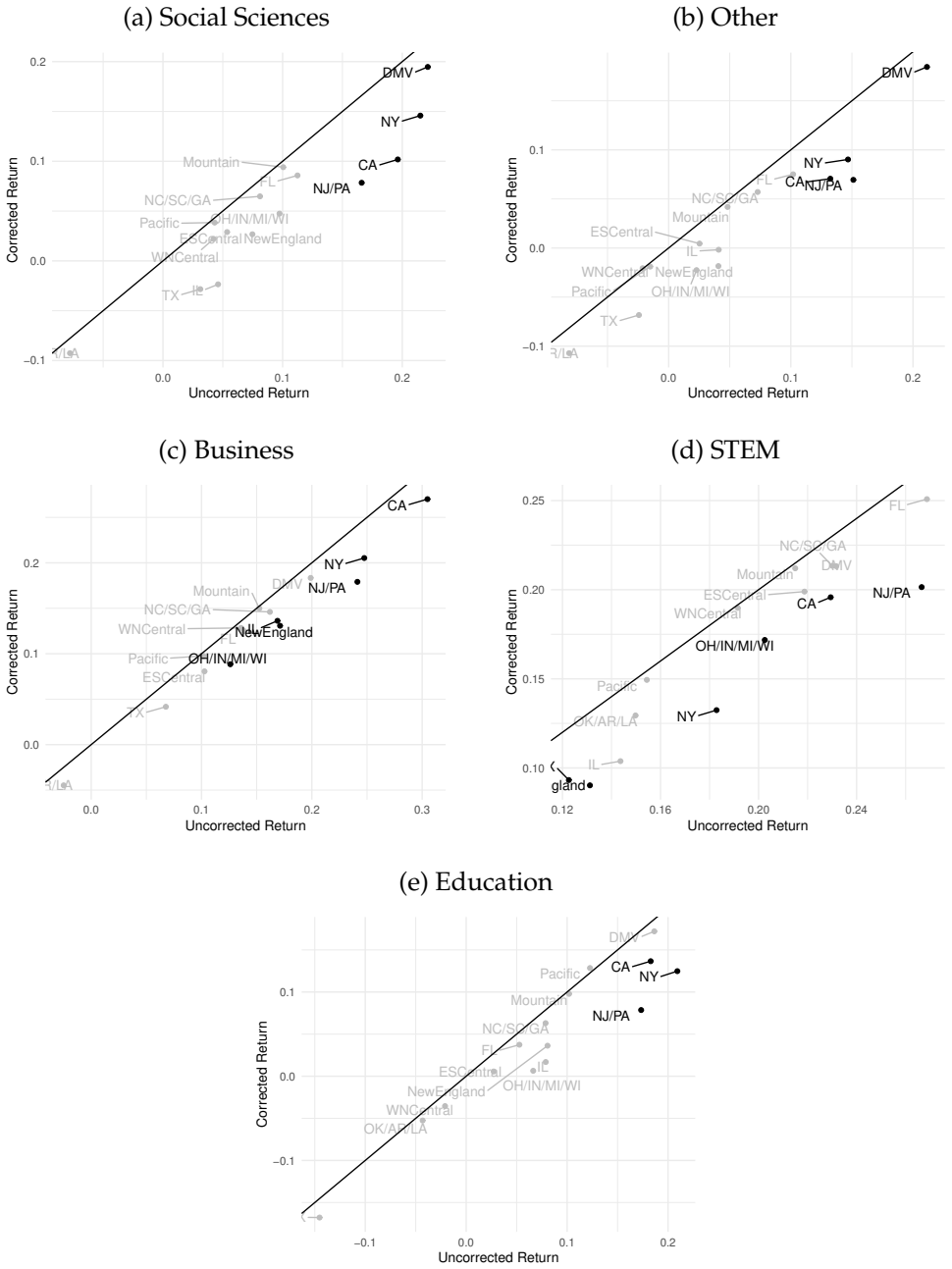
Figure F5: Scatter plots of uncorrected and corrected returns to major and working in an unrelated occupation, adv. degree holders



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Dots mark the uncorrected and corrected returns. Gray colored dots and labels indicate statistically and or economically insignificant differences at the 5% level. See Appendix D for further details.

Source: Author's calculations from American Community Survey, 2010-2019.

Figure F6: Scatter plots of uncorrected and corrected returns to major and working in a related occupation, adv. degree holders



Notes: Scatter plots of return to major for those working in an unrelated occupation. Solid black lines are 45-degree lines. Dots mark the uncorrected and corrected returns. Gray colored dots and labels indicate statistically and or economically insignificant differences at the 5% level. See Appendix D for further details.

Source: Author's calculations from American Community Survey, 2010-2019.

Table F10: Return to Bach. Deg. Social Science majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Social Science Return	Corrected Social Science Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.048	0.084	105.5	18.930	18.189
	(0.033)	(0.034)	(1006.9)	[0.000]	[0.000]
E S Central Div	0.023	0.010	-34.5	5.223	8.436
	(0.025)	(0.025)	(625.1)	[0.022]	[0.000]
Florida	0.060	0.062	0.7	0.056	6.034
	(0.033)	(0.034)	(52.2)	[0.813]	[0.000]
Illinois	0.060	0.068	-4.8	0.252	21.029
	(0.030)	(0.034)	(117.6)	[0.615]	[0.000]
Mountain Div	0.064	0.069	11.1	2.423	4.502
	(0.025)	(0.025)	(47.5)	[0.120]	[0.000]
NC, SC, GA	0.102	0.096	-3.0	1.393	9.167
	(0.023)	(0.023)	(5.5)	[0.238]	[0.000]
New England	0.125	0.108	-10.3	4.347	18.820
	(0.027)	(0.029)	(8.3)	[0.037]	[0.000]
New Jersey and Penn.	0.113	0.116	3.2	0.046	23.867
	(0.023)	(0.026)	(11.5)	[0.830]	[0.000]
New York	0.105	0.071	-44.1	7.500	29.037
	(0.030)	(0.032)	(21.8)	[0.006]	[0.000]
OH, IN, MI, WI	0.015	0.027	45.2	2.046	24.684
	(0.017)	(0.019)	(1001.7)	[0.153]	[0.000]
OK, AR, LA	-0.028	-0.031	-0.1	0.082	3.409
	(0.036)	(0.037)	(238.6)	[0.775]	[0.000]
OR, WA, AK, HI	0.051	0.049	-18.0	0.080	3.561
	(0.029)	(0.030)	(273.4)	[0.778]	[0.000]
Texas	0.054	0.053	-9.5	0.008	7.244
	(0.031)	(0.033)	(175.4)	[0.929]	[0.000]
W N Central Div	0.011	0.019	-43.1	0.807	10.384
	(0.020)	(0.022)	(2653.8)	[0.369]	[0.000]
WV, VA, DC, MD, DE	0.067	0.079	18.7	5.611	16.208
	(0.028)	(0.029)	(58.4)	[0.018]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F11: Return to Bach. Deg. Social Science majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Social Science Return	Corrected Social Science Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.317	0.368	14.5	22.067	39.768
	(0.023)	(0.026)	(3.5)	[0.000]	[0.000]
E S Central Div	0.188	0.181	-1.3	0.538	6.541
	(0.025)	(0.026)	(4.5)	[0.463]	[0.000]
Florida	0.365	0.377	2.3	2.279	9.504
	(0.023)	(0.024)	(1.7)	[0.131]	[0.000]
Illinois	0.293	0.290	-4.6	0.064	28.640
	(0.029)	(0.032)	(5.4)	[0.800]	[0.000]
Mountain Div	0.328	0.331	0.5	0.488	6.660
	(0.018)	(0.019)	(1.6)	[0.485]	[0.000]
NC, SC, GA	0.301	0.299	-0.2	0.080	14.025
	(0.019)	(0.020)	(2.4)	[0.777]	[0.000]
New England	0.276	0.260	-3.8	1.836	34.959
	(0.023)	(0.026)	(4.0)	[0.175]	[0.000]
New Jersey and Penn.	0.188	0.199	4.3	0.664	39.428
	(0.019)	(0.023)	(6.9)	[0.415]	[0.000]
New York	0.241	0.222	-13.3	2.122	47.313
	(0.033)	(0.036)	(6.2)	[0.145]	[0.000]
OH, IN, MI, WI	0.183	0.204	6.0	3.842	42.131
	(0.016)	(0.019)	(5.2)	[0.050]	[0.000]
OK, AR, LA	0.246	0.249	-1.3	0.123	4.082
	(0.034)	(0.035)	(3.7)	[0.725]	[0.000]
OR, WA, AK, HI	0.258	0.269	3.1	1.171	10.790
	(0.024)	(0.026)	(3.3)	[0.279]	[0.000]
Texas	0.192	0.218	11.0	4.286	18.653
	(0.023)	(0.026)	(6.2)	[0.038]	[0.000]
W N Central Div	0.185	0.194	2.0	1.095	17.671
	(0.020)	(0.022)	(4.9)	[0.295]	[0.000]
WV, VA, DC, MD, DE	0.336	0.345	2.0	1.683	20.536
	(0.018)	(0.020)	(1.9)	[0.195]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F12: Return to Bach. Deg. Other majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Other Return	Corrected Other Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	-0.002	0.014	145.9	14.603	18.189
	(0.032)	(0.032)	(4673.9)	[0.000]	[0.000]
E S Central Div	0.046	0.034	-18.3	3.516	8.436
	(0.022)	(0.023)	(127.8)	[0.061]	[0.000]
Florida	0.075	0.072	-7.9	0.233	6.034
	(0.031)	(0.032)	(27.9)	[0.629]	[0.000]
Illinois	0.020	-0.009	118.5	12.261	21.029
	(0.027)	(0.029)	(4679.7)	[0.000]	[0.000]
Mountain Div	0.069	0.068	3.3	0.010	4.502
	(0.025)	(0.025)	(31.6)	[0.919]	[0.000]
NC, SC, GA	0.081	0.073	-7.2	2.627	9.167
	(0.022)	(0.023)	(10.8)	[0.105]	[0.000]
New England	0.069	0.053	-30.5	3.923	18.820
	(0.025)	(0.027)	(76.0)	[0.048]	[0.000]
New Jersey and Penn.	0.079	0.074	-8.2	0.131	23.867
	(0.020)	(0.024)	(34.8)	[0.717]	[0.000]
New York	0.029	0.005	-10.6	3.448	29.037
	(0.027)	(0.030)	(5908.6)	[0.063]	[0.000]
OH, IN, MI, WI	0.016	0.016	23.4	0.003	24.684
	(0.017)	(0.018)	(730.4)	[0.958]	[0.000]
OK, AR, LA	0.029	0.027	-37.4	0.052	3.409
	(0.035)	(0.036)	(291.0)	[0.820]	[0.000]
OR, WA, AK, HI	0.050	0.047	-4.8	0.430	3.561
	(0.028)	(0.029)	(128.3)	[0.512]	[0.000]
Texas	0.059	0.048	-2.5	1.631	7.244
	(0.029)	(0.030)	(268.3)	[0.202]	[0.000]
W N Central Div	0.022	0.013	0.9	1.060	10.384
	(0.019)	(0.021)	(448.5)	[0.303]	[0.000]
WV, VA, DC, MD, DE	0.051	0.057	6.4	4.663	16.208
	(0.028)	(0.028)	(78.0)	[0.031]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F13: Return to Bach. Deg. Other majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Other Return	Corrected Other Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.233	0.252	10.3	7.662	39.768
	(0.021)	(0.023)	(2.8)	[0.006]	[0.000]
E S Central Div	0.218	0.209	-1.8	1.178	6.541
	(0.019)	(0.020)	(3.8)	[0.278]	[0.000]
Florida	0.309	0.316	0.5	1.090	9.504
	(0.017)	(0.019)	(1.8)	[0.296]	[0.000]
Illinois	0.218	0.186	-12.1	10.637	28.640
	(0.020)	(0.022)	(4.6)	[0.001]	[0.000]
Mountain Div	0.316	0.317	0.1	0.004	6.660
	(0.015)	(0.016)	(1.7)	[0.948]	[0.000]
NC, SC, GA	0.253	0.250	-1.0	0.132	14.025
	(0.015)	(0.017)	(2.7)	[0.717]	[0.000]
New England	0.207	0.201	-3.7	0.287	34.959
	(0.019)	(0.023)	(5.0)	[0.592]	[0.000]
New Jersey and Penn.	0.155	0.161	1.8	0.309	39.428
	(0.013)	(0.018)	(8.2)	[0.578]	[0.000]
New York	0.184	0.161	-21.5	3.763	47.313
	(0.030)	(0.032)	(8.5)	[0.052]	[0.000]
OH, IN, MI, WI	0.207	0.226	3.9	4.069	42.131
	(0.011)	(0.015)	(4.1)	[0.044]	[0.000]
OK, AR, LA	0.252	0.265	1.8	1.635	4.082
	(0.025)	(0.027)	(4.4)	[0.201]	[0.000]
OR, WA, AK, HI	0.261	0.269	1.3	0.498	10.790
	(0.022)	(0.025)	(3.3)	[0.480]	[0.000]
Texas	0.167	0.179	8.3	1.842	18.653
	(0.016)	(0.018)	(5.3)	[0.175]	[0.000]
W N Central Div	0.192	0.183	-2.0	1.716	17.671
	(0.014)	(0.015)	(4.3)	[0.190]	[0.000]
WV, VA, DC, MD, DE	0.255	0.267	4.3	4.898	20.536
	(0.016)	(0.016)	(2.4)	[0.027]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F14: Return to Bach. Deg. Business majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Business Return	Corrected Business Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.142	0.143	0.2	0.028	18.189
	(0.032)	(0.032)	(4.3)	[0.866]	[0.000]
E S Central Div	0.189	0.177	-5.0	2.160	8.436
	(0.022)	(0.024)	(3.5)	[0.142]	[0.000]
Florida	0.202	0.195	-2.0	1.203	6.034
	(0.030)	(0.031)	(2.4)	[0.273]	[0.000]
Illinois	0.214	0.188	-11.5	12.997	21.029
	(0.027)	(0.028)	(4.6)	[0.000]	[0.000]
Mountain Div	0.206	0.198	-2.7	9.169	4.502
	(0.023)	(0.024)	(2.6)	[0.002]	[0.000]
NC, SC, GA	0.204	0.195	-3.1	4.239	9.167
	(0.022)	(0.023)	(2.5)	[0.039]	[0.000]
New England	0.243	0.207	-10.8	17.713	18.820
	(0.026)	(0.028)	(4.1)	[0.000]	[0.000]
New Jersey and Penn.	0.248	0.223	-8.6	11.506	23.867
	(0.021)	(0.022)	(2.8)	[0.001]	[0.000]
New York	0.218	0.158	-35.6	14.387	29.037
	(0.027)	(0.031)	(8.7)	[0.000]	[0.000]
OH, IN, MI, WI	0.183	0.165	-12.6	12.433	24.684
	(0.016)	(0.017)	(3.1)	[0.000]	[0.000]
OK, AR, LA	0.170	0.163	-5.0	1.035	3.409
	(0.035)	(0.035)	(4.1)	[0.309]	[0.000]
OR, WA, AK, HI	0.220	0.216	-2.1	0.522	3.561
	(0.029)	(0.030)	(2.4)	[0.470]	[0.000]
Texas	0.195	0.190	-1.4	0.548	7.244
	(0.028)	(0.029)	(2.8)	[0.459]	[0.000]
W N Central Div	0.194	0.183	-7.9	5.123	10.384
	(0.018)	(0.018)	(3.0)	[0.024]	[0.000]
WV, VA, DC, MD, DE	0.179	0.174	-4.1	1.358	16.208
	(0.028)	(0.028)	(2.8)	[0.244]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F15: Return to Bach. Deg. Business majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Business Return	Corrected Business Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.466	0.455	-0.7	3.890	39.768
	(0.022)	(0.022)	(1.2)	[0.049]	[0.000]
E S Central Div	0.463	0.456	-0.8	0.859	6.541
	(0.017)	(0.019)	(1.4)	[0.354]	[0.000]
Florida	0.537	0.533	0.1	0.628	9.504
	(0.016)	(0.016)	(0.8)	[0.428]	[0.000]
Illinois	0.485	0.449	-7.2	43.685	28.640
	(0.020)	(0.021)	(1.5)	[0.000]	[0.000]
Mountain Div	0.524	0.516	-1.2	3.525	6.660
	(0.014)	(0.015)	(0.8)	[0.060]	[0.000]
NC, SC, GA	0.520	0.521	-0.1	0.022	14.025
	(0.014)	(0.016)	(1.0)	[0.881]	[0.000]
New England	0.464	0.438	-4.6	8.187	34.959
	(0.019)	(0.021)	(1.7)	[0.004]	[0.000]
New Jersey and Penn.	0.423	0.398	-5.5	16.530	39.428
	(0.012)	(0.013)	(1.6)	[0.000]	[0.000]
New York	0.478	0.423	-13.0	15.439	47.313
	(0.029)	(0.032)	(3.2)	[0.000]	[0.000]
OH, IN, MI, WI	0.467	0.465	-2.6	0.165	42.131
	(0.011)	(0.013)	(1.4)	[0.684]	[0.000]
OK, AR, LA	0.533	0.540	0.2	0.925	4.082
	(0.023)	(0.024)	(1.5)	[0.336]	[0.000]
OR, WA, AK, HI	0.485	0.491	0.5	0.720	10.790
	(0.020)	(0.022)	(1.0)	[0.396]	[0.000]
Texas	0.430	0.429	0.8	0.019	18.653
	(0.015)	(0.016)	(1.3)	[0.891]	[0.000]
W N Central Div	0.432	0.425	-2.0	1.680	17.671
	(0.013)	(0.014)	(1.4)	[0.195]	[0.000]
WV, VA, DC, MD, DE	0.516	0.514	-0.1	0.285	20.536
	(0.016)	(0.017)	(0.8)	[0.594]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F16: Return to Bach. Deg. STEM majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected STEM Return	Corrected STEM Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.163	0.162	-2.2	8.676	18.189
	(0.032)	(0.032)	(3.7)	[0.003]	[0.000]
E S Central Div	0.249	0.235	-5.1	3.773	8.436
	(0.022)	(0.024)	(2.7)	[0.052]	[0.000]
Florida	0.205	0.195	-3.9	1.899	6.034
	(0.031)	(0.031)	(2.5)	[0.168]	[0.000]
Illinois	0.222	0.187	-14.8	44.547	21.029
	(0.028)	(0.029)	(4.7)	[0.000]	[0.000]
Mountain Div	0.241	0.233	-2.6	11.620	4.502
	(0.024)	(0.024)	(2.3)	[0.001]	[0.000]
NC, SC, GA	0.268	0.259	-2.7	4.459	9.167
	(0.022)	(0.022)	(1.9)	[0.035]	[0.000]
New England	0.264	0.224	-11.6	21.786	18.820
	(0.027)	(0.028)	(3.8)	[0.000]	[0.000]
New Jersey and Penn.	0.301	0.272	-9.2	12.491	23.867
	(0.021)	(0.022)	(2.7)	[0.000]	[0.000]
New York	0.203	0.138	-39.6	16.806	29.037
	(0.027)	(0.032)	(9.5)	[0.000]	[0.000]
OH, IN, MI, WI	0.216	0.191	-13.9	16.616	24.684
	(0.016)	(0.017)	(2.9)	[0.000]	[0.000]
OK, AR, LA	0.238	0.231	-4.4	0.794	3.409
	(0.032)	(0.033)	(3.2)	[0.373]	[0.000]
OR, WA, AK, HI	0.228	0.223	-2.2	0.544	3.561
	(0.029)	(0.030)	(2.7)	[0.461]	[0.000]
Texas	0.275	0.268	-1.8	1.139	7.244
	(0.028)	(0.029)	(2.1)	[0.286]	[0.000]
W N Central Div	0.205	0.187	-9.7	7.722	10.384
	(0.018)	(0.019)	(3.2)	[0.005]	[0.000]
WV, VA, DC, MD, DE	0.230	0.222	-4.6	16.525	16.208
	(0.027)	(0.027)	(2.3)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F17: Return to Bach. Deg. STEM majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected STEM Return	Corrected STEM Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.468	0.454	-2.1	7.359	39.768
	(0.021)	(0.022)	(1.2)	[0.007]	[0.000]
E S Central Div	0.453	0.441	-1.8	3.420	6.541
	(0.017)	(0.018)	(1.5)	[0.064]	[0.000]
Florida	0.492	0.488	-1.0	0.775	9.504
	(0.016)	(0.016)	(0.9)	[0.379]	[0.000]
Illinois	0.431	0.387	-10.1	29.155	28.640
	(0.019)	(0.021)	(2.0)	[0.000]	[0.000]
Mountain Div	0.524	0.516	-1.6	4.223	6.660
	(0.013)	(0.014)	(0.8)	[0.040]	[0.000]
NC, SC, GA	0.487	0.486	-0.6	0.066	14.025
	(0.014)	(0.015)	(1.2)	[0.797]	[0.000]
New England	0.420	0.386	-6.6	14.633	34.959
	(0.018)	(0.020)	(2.0)	[0.000]	[0.000]
New Jersey and Penn.	0.372	0.349	-6.9	7.606	39.428
	(0.012)	(0.015)	(2.3)	[0.006]	[0.000]
New York	0.416	0.358	-16.1	16.351	47.313
	(0.028)	(0.032)	(3.8)	[0.000]	[0.000]
OH, IN, MI, WI	0.440	0.430	-4.5	2.137	42.131
	(0.010)	(0.012)	(1.4)	[0.144]	[0.000]
OK, AR, LA	0.504	0.512	0.2	0.775	4.082
	(0.022)	(0.024)	(1.9)	[0.379]	[0.000]
OR, WA, AK, HI	0.511	0.513	-0.2	0.105	10.790
	(0.019)	(0.021)	(1.0)	[0.746]	[0.000]
Texas	0.385	0.383	1.1	0.076	18.653
	(0.015)	(0.016)	(1.7)	[0.783]	[0.000]
W N Central Div	0.417	0.400	-3.8	9.121	17.671
	(0.012)	(0.013)	(1.5)	[0.003]	[0.000]
WV, VA, DC, MD, DE	0.488	0.484	-0.8	1.240	20.536
	(0.014)	(0.015)	(0.9)	[0.265]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F18: Return to Adv. Deg. Education majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Education Return	Corrected Education Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.143	0.112	-5.3	25.460	18.189
	(0.065)	(0.064)	(434.4)	[0.000]	[0.000]
E S Central Div	0.241	0.243	-1.6	0.079	8.436
	(0.065)	(0.065)	(5.2)	[0.779]	[0.000]
Florida	0.185	0.177	-3.1	2.579	6.034
	(0.077)	(0.077)	(20.9)	[0.108]	[0.000]
Illinois	0.153	0.069	-67.8	9.802	21.029
	(0.077)	(0.082)	(724.1)	[0.002]	[0.000]
Mountain Div	0.299	0.303	0.8	0.172	4.502
	(0.063)	(0.064)	(2.5)	[0.678]	[0.000]
NC, SC, GA	0.157	0.152	-1.7	2.778	9.167
	(0.058)	(0.058)	(13.2)	[0.096]	[0.000]
New England	0.111	0.054	-48.6	20.583	18.820
	(0.060)	(0.061)	(676.7)	[0.000]	[0.000]
New Jersey and Penn.	0.221	0.145	-35.9	12.578	23.867
	(0.054)	(0.059)	(17.3)	[0.000]	[0.000]
New York	0.193	0.079	-80.2	8.146	29.037
	(0.058)	(0.071)	(43.4)	[0.004]	[0.000]
OH, IN, MI, WI	0.199	0.135	-32.9	8.245	24.684
	(0.047)	(0.052)	(15.3)	[0.004]	[0.000]
OK, AR, LA	0.154	0.157	3.2	0.324	3.409
	(0.103)	(0.103)	(91.6)	[0.569]	[0.000]
OR, WA, AK, HI	0.258	0.251	-3.8	0.589	3.561
	(0.082)	(0.082)	(5.0)	[0.443]	[0.000]
Texas	0.034	0.023	6.4	10.754	7.244
	(0.061)	(0.061)	(241.4)	[0.001]	[0.000]
W N Central Div	0.072	0.050	-6.7	3.548	10.384
	(0.061)	(0.062)	(191.9)	[0.060]	[0.000]
WV, VA, DC, MD, DE	0.175	0.168	-6.0	3.924	16.208
	(0.055)	(0.055)	(5.7)	[0.048]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F19: Return to Adv. Deg. Education majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Education Return	Corrected Education Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.183	0.137	-26.1	25.264	39.768
	(0.038)	(0.039)	(8.7)	[0.000]	[0.000]
E S Central Div	0.027	0.006	34.8	2.558	6.541
	(0.040)	(0.043)	(881.0)	[0.110]	[0.000]
Florida	0.053	0.037	-167.4	1.634	9.504
	(0.047)	(0.049)	(3099.7)	[0.201]	[0.000]
Illinois	0.079	0.017	-33.9	14.006	28.640
	(0.047)	(0.050)	(689.4)	[0.000]	[0.000]
Mountain Div	0.102	0.098	-26.2	1.382	6.660
	(0.039)	(0.039)	(387.9)	[0.240]	[0.000]
NC, SC, GA	0.079	0.063	-23.1	3.473	14.025
	(0.034)	(0.035)	(42.2)	[0.062]	[0.000]
New England	0.081	0.036	-136.4	64.535	34.959
	(0.036)	(0.037)	(1710.5)	[0.000]	[0.000]
New Jersey and Penn.	0.173	0.079	-40.6	19.106	39.428
	(0.032)	(0.038)	(15.3)	[0.000]	[0.000]
New York	0.209	0.125	-41.2	7.596	47.313
	(0.041)	(0.051)	(17.3)	[0.006]	[0.000]
OH, IN, MI, WI	0.066	0.006	711.9	15.705	42.131
	(0.030)	(0.034)	(18192.7)	[0.000]	[0.000]
OK, AR, LA	-0.043	-0.053	30.2	0.264	4.082
	(0.059)	(0.061)	(861.8)	[0.608]	[0.000]
OR, WA, AK, HI	0.123	0.128	1.5	0.321	10.790
	(0.053)	(0.054)	(18.5)	[0.571]	[0.000]
Texas	-0.145	-0.168	13.1	28.495	18.653
	(0.039)	(0.039)	(10.0)	[0.000]	[0.000]
W N Central Div	-0.021	-0.035	48.6	1.192	17.671
	(0.035)	(0.038)	(1250.8)	[0.275]	[0.000]
WV, VA, DC, MD, DE	0.186	0.172	-7.1	6.838	20.536
	(0.035)	(0.035)	(3.8)	[0.009]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F20: Return to Adv. Deg. Social Science majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Social Science Return	Corrected Social Science Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.246	0.183	-24.5	26.482	18.189
	(0.043)	(0.045)	(6.8)	[0.000]	[0.000]
E S Central Div	0.253	0.251	-3.2	0.674	8.436
	(0.062)	(0.062)	(4.0)	[0.412]	[0.000]
Florida	0.260	0.245	-3.5	14.707	6.034
	(0.066)	(0.066)	(4.6)	[0.000]	[0.000]
Illinois	0.263	0.180	-25.5	24.677	21.029
	(0.066)	(0.068)	(9.8)	[0.000]	[0.000]
Mountain Div	0.305	0.305	0.0	0.002	4.502
	(0.052)	(0.053)	(2.1)	[0.969]	[0.000]
NC, SC, GA	0.210	0.202	-3.5	2.076	9.167
	(0.047)	(0.048)	(3.0)	[0.150]	[0.000]
New England	0.199	0.146	-26.9	35.622	18.820
	(0.048)	(0.049)	(9.9)	[0.000]	[0.000]
New Jersey and Penn.	0.226	0.155	-31.9	38.373	23.867
	(0.046)	(0.047)	(9.2)	[0.000]	[0.000]
New York	0.272	0.190	-34.3	25.420	29.037
	(0.049)	(0.052)	(10.1)	[0.000]	[0.000]
OH, IN, MI, WI	0.277	0.240	-11.6	10.624	24.684
	(0.046)	(0.047)	(4.1)	[0.001]	[0.000]
OK, AR, LA	0.162	0.158	-3.4	0.398	3.409
	(0.101)	(0.101)	(32.4)	[0.528]	[0.000]
OR, WA, AK, HI	0.349	0.339	-3.4	1.022	3.561
	(0.071)	(0.072)	(2.8)	[0.312]	[0.000]
Texas	0.094	0.079	-10.1	2.678	7.244
	(0.051)	(0.052)	(83.9)	[0.102]	[0.000]
W N Central Div	0.138	0.111	-12.7	16.379	10.384
	(0.053)	(0.054)	(28.8)	[0.000]	[0.000]
WV, VA, DC, MD, DE	0.280	0.265	-6.4	1076.137	16.208
	(0.038)	(0.038)	(2.4)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F21: Return to Adv. Deg. Social Science majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Social Science Return	Corrected Social Science Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.196	0.102	-44.8	69.395	39.768
	(0.033)	(0.035)	(10.7)	[0.000]	[0.000]
E S Central Div	0.054	0.029	-56.0	4.171	6.541
	(0.046)	(0.047)	(409.6)	[0.041]	[0.000]
Florida	0.112	0.086	-32.3	4.517	9.504
	(0.046)	(0.048)	(125.4)	[0.034]	[0.000]
Illinois	0.046	-0.024	-3.6	19.407	28.640
	(0.049)	(0.051)	(1752.0)	[0.000]	[0.000]
Mountain Div	0.101	0.094	-9.3	1.695	6.660
	(0.041)	(0.041)	(22.1)	[0.193]	[0.000]
NC, SC, GA	0.081	0.065	-23.7	6.687	14.025
	(0.038)	(0.039)	(79.2)	[0.010]	[0.000]
New England	0.075	0.027	-120.6	49.124	34.959
	(0.037)	(0.037)	(570.5)	[0.000]	[0.000]
New Jersey and Penn.	0.166	0.078	-48.8	68.647	39.428
	(0.034)	(0.035)	(13.5)	[0.000]	[0.000]
New York	0.215	0.146	-31.5	22.221	47.313
	(0.039)	(0.041)	(10.6)	[0.000]	[0.000]
OH, IN, MI, WI	0.097	0.047	-60.2	22.462	42.131
	(0.033)	(0.034)	(85.7)	[0.000]	[0.000]
OK, AR, LA	-0.078	-0.093	2.2	1.059	4.082
	(0.065)	(0.067)	(569.3)	[0.303]	[0.000]
OR, WA, AK, HI	0.043	0.038	-61.7	0.619	10.790
	(0.050)	(0.050)	(1337.3)	[0.431]	[0.000]
Texas	0.031	-0.028	62.9	15.975	18.653
	(0.043)	(0.046)	(2238.7)	[0.000]	[0.000]
W N Central Div	0.042	0.022	9.1	3.887	17.671
	(0.041)	(0.042)	(807.0)	[0.049]	[0.000]
WV, VA, DC, MD, DE	0.221	0.195	-10.6	24.800	20.536
	(0.033)	(0.033)	(3.3)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F22: Return to Adv. Deg. Other majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Other Return	Corrected Other Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.206	0.165	-21.1	26.033	18.189
	(0.043)	(0.043)	(6.7)	[0.000]	[0.000]
E S Central Div	0.217	0.215	-3.3	0.093	8.436
	(0.059)	(0.058)	(4.8)	[0.761]	[0.000]
Florida	0.167	0.155	71.8	4.119	6.034
	(0.065)	(0.065)	(1709.5)	[0.042]	[0.000]
Illinois	0.268	0.222	-16.0	9.308	21.029
	(0.063)	(0.065)	(6.9)	[0.002]	[0.000]
Mountain Div	0.284	0.285	0.4	0.018	4.502
	(0.052)	(0.052)	(2.3)	[0.893]	[0.000]
NC, SC, GA	0.129	0.124	-4.1	19.072	9.167
	(0.048)	(0.048)	(7.0)	[0.000]	[0.000]
New England	0.153	0.100	-35.0	38.949	18.820
	(0.045)	(0.046)	(22.1)	[0.000]	[0.000]
New Jersey and Penn.	0.222	0.161	-28.2	38.634	23.867
	(0.045)	(0.046)	(8.5)	[0.000]	[0.000]
New York	0.230	0.145	-39.0	35.366	29.037
	(0.043)	(0.046)	(12.1)	[0.000]	[0.000]
OH, IN, MI, WI	0.236	0.214	-8.6	5.602	24.684
	(0.045)	(0.046)	(4.1)	[0.018]	[0.000]
OK, AR, LA	0.108	0.105	-10.6	19.148	3.409
	(0.091)	(0.091)	(88.3)	[0.000]	[0.000]
OR, WA, AK, HI	0.231	0.224	-4.2	0.392	3.561
	(0.069)	(0.070)	(5.1)	[0.531]	[0.000]
Texas	0.014	0.009	-10.2	1.688	7.244
	(0.049)	(0.049)	(472.3)	[0.194]	[0.000]
W N Central Div	0.080	0.071	-16.4	0.984	10.384
	(0.054)	(0.055)	(121.4)	[0.321]	[0.000]
WV, VA, DC, MD, DE	0.228	0.215	-6.8	37.429	16.208
	(0.040)	(0.040)	(3.1)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F23: Return to Adv. Deg. Other majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Other Return	Corrected Other Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.132	0.071	-49.8	82.075	39.768
	(0.031)	(0.032)	(15.1)	[0.000]	[0.000]
E S Central Div	0.025	0.004	200.9	2.802	6.541
	(0.038)	(0.040)	(4513.3)	[0.094]	[0.000]
Florida	0.102	0.075	-36.1	8.695	9.504
	(0.046)	(0.047)	(118.5)	[0.003]	[0.000]
Illinois	0.041	-0.002	-113.3	16.771	28.640
	(0.046)	(0.047)	(1098.6)	[0.000]	[0.000]
Mountain Div	0.048	0.042	-24.8	2.620	6.660
	(0.038)	(0.037)	(384.6)	[0.106]	[0.000]
NC, SC, GA	0.073	0.057	-37.7	9.070	14.025
	(0.035)	(0.035)	(256.3)	[0.003]	[0.000]
New England	0.041	-0.018	-61.4	81.598	34.959
	(0.033)	(0.034)	(1132.3)	[0.000]	[0.000]
New Jersey and Penn.	0.151	0.069	-49.7	81.074	39.428
	(0.032)	(0.033)	(15.5)	[0.000]	[0.000]
New York	0.147	0.090	-35.9	13.638	47.313
	(0.033)	(0.037)	(14.8)	[0.000]	[0.000]
OH, IN, MI, WI	0.023	-0.023	-118.8	37.232	42.131
	(0.030)	(0.031)	(3747.2)	[0.000]	[0.000]
OK, AR, LA	-0.081	-0.107	148.8	4.540	4.082
	(0.061)	(0.062)	(1736.5)	[0.033]	[0.000]
OR, WA, AK, HI	-0.015	-0.019	-18.8	0.348	10.790
	(0.051)	(0.052)	(228.1)	[0.555]	[0.000]
Texas	-0.024	-0.068	110.2	23.684	18.653
	(0.036)	(0.037)	(2320.6)	[0.000]	[0.000]
W N Central Div	-0.021	-0.021	-244.9	0.006	17.671
	(0.037)	(0.038)	(4907.3)	[0.941]	[0.000]
WV, VA, DC, MD, DE	0.211	0.184	-12.8	24.032	20.536
	(0.031)	(0.031)	(3.5)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F24: Return to Adv. Deg. Business majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected Business Return	Corrected Business Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.215	0.184	-14.2	30.252	18.189
	(0.042)	(0.043)	(4.6)	[0.000]	[0.000]
E S Central Div	0.230	0.231	-1.4	0.019	8.436
	(0.058)	(0.057)	(3.8)	[0.891]	[0.000]
Florida	0.150	0.140	-7.7	3.409	6.034
	(0.063)	(0.063)	(41.1)	[0.065]	[0.000]
Illinois	0.291	0.248	-12.8	9.711	21.029
	(0.062)	(0.063)	(6.1)	[0.002]	[0.000]
Mountain Div	0.265	0.269	1.2	0.241	4.502
	(0.050)	(0.050)	(2.5)	[0.624]	[0.000]
NC, SC, GA	0.187	0.180	-2.6	21.409	9.167
	(0.045)	(0.045)	(3.6)	[0.000]	[0.000]
New England	0.147	0.109	-30.9	33.899	18.820
	(0.047)	(0.048)	(25.3)	[0.000]	[0.000]
New Jersey and Penn.	0.244	0.190	-22.6	35.503	23.867
	(0.045)	(0.046)	(7.0)	[0.000]	[0.000]
New York	0.260	0.207	-25.2	16.373	29.037
	(0.048)	(0.049)	(8.3)	[0.000]	[0.000]
OH, IN, MI, WI	0.289	0.265	-7.0	4.471	24.684
	(0.046)	(0.048)	(3.1)	[0.034]	[0.000]
OK, AR, LA	0.161	0.160	-2.4	0.002	3.409
	(0.090)	(0.090)	(50.9)	[0.964]	[0.000]
OR, WA, AK, HI	0.313	0.304	-3.2	1.398	3.561
	(0.069)	(0.069)	(3.2)	[0.237]	[0.000]
Texas	0.056	0.057	0.8	0.085	7.244
	(0.049)	(0.049)	(83.3)	[0.770]	[0.000]
W N Central Div	0.106	0.090	-10.8	7.533	10.384
	(0.054)	(0.054)	(45.9)	[0.006]	[0.000]
WV, VA, DC, MD, DE	0.245	0.237	-3.6	62.531	16.208
	(0.041)	(0.041)	(2.5)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F25: Return to Adv. Deg. Business majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected Business Return	Corrected Business Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.305	0.270	-13.1	55.266	39.768
	(0.032)	(0.032)	(2.8)	[0.000]	[0.000]
E S Central Div	0.103	0.081	-24.4	5.411	6.541
	(0.037)	(0.039)	(45.5)	[0.020]	[0.000]
Florida	0.139	0.125	-18.4	2.176	9.504
	(0.041)	(0.042)	(18.7)	[0.140]	[0.000]
Illinois	0.169	0.136	-19.2	11.101	28.640
	(0.040)	(0.041)	(8.4)	[0.001]	[0.000]
Mountain Div	0.152	0.150	-3.5	0.425	6.660
	(0.037)	(0.037)	(4.2)	[0.514]	[0.000]
NC, SC, GA	0.162	0.146	-9.6	5.804	14.025
	(0.033)	(0.033)	(5.2)	[0.016]	[0.000]
New England	0.171	0.131	-25.5	39.745	34.959
	(0.030)	(0.030)	(6.8)	[0.000]	[0.000]
New Jersey and Penn.	0.241	0.179	-21.4	50.356	39.428
	(0.030)	(0.031)	(4.9)	[0.000]	[0.000]
New York	0.248	0.205	-16.1	8.593	47.313
	(0.031)	(0.034)	(6.6)	[0.003]	[0.000]
OH, IN, MI, WI	0.126	0.088	-29.8	33.867	42.131
	(0.029)	(0.030)	(12.8)	[0.000]	[0.000]
OK, AR, LA	-0.025	-0.045	18.7	2.283	4.082
	(0.059)	(0.061)	(1145.4)	[0.131]	[0.000]
OR, WA, AK, HI	0.102	0.098	-16.4	0.771	10.790
	(0.049)	(0.050)	(215.7)	[0.380]	[0.000]
Texas	0.068	0.042	-30.9	11.147	18.653
	(0.034)	(0.035)	(256.7)	[0.001]	[0.000]
W N Central Div	0.136	0.128	-7.1	0.430	17.671
	(0.035)	(0.037)	(9.4)	[0.512]	[0.000]
WV, VA, DC, MD, DE	0.199	0.183	-7.7	6.948	20.536
	(0.030)	(0.031)	(3.2)	[0.008]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F26: Return to Adv. Deg. STEM majors in Unrelated occupation, by state (uncorrected and corrected)

Location	Uncorrected STEM Return	Corrected STEM Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.310	0.279	-9.0	76.225	18.189
	(0.040)	(0.040)	(2.5)	[0.000]	[0.000]
E S Central Div	0.277	0.273	-3.5	1.649	8.436
	(0.056)	(0.056)	(3.4)	[0.199]	[0.000]
Florida	0.231	0.221	-2.8	1.768	6.034
	(0.060)	(0.059)	(4.2)	[0.184]	[0.000]
Illinois	0.314	0.274	-11.9	19.482	21.029
	(0.059)	(0.060)	(4.7)	[0.000]	[0.000]
Mountain Div	0.322	0.324	0.7	0.120	4.502
	(0.052)	(0.052)	(2.2)	[0.729]	[0.000]
NC, SC, GA	0.184	0.171	-5.1	10.318	9.167
	(0.045)	(0.045)	(3.7)	[0.001]	[0.000]
New England	0.200	0.160	-21.8	18.350	18.820
	(0.044)	(0.045)	(7.8)	[0.000]	[0.000]
New Jersey and Penn.	0.245	0.191	-21.7	48.780	23.867
	(0.044)	(0.044)	(6.3)	[0.000]	[0.000]
New York	0.350	0.291	-19.6	28.007	29.037
	(0.045)	(0.047)	(5.0)	[0.000]	[0.000]
OH, IN, MI, WI	0.336	0.317	-5.2	3.483	24.684
	(0.042)	(0.044)	(2.4)	[0.062]	[0.000]
OK, AR, LA	0.202	0.199	-3.7	0.148	3.409
	(0.090)	(0.090)	(13.5)	[0.701]	[0.000]
OR, WA, AK, HI	0.358	0.351	-2.1	0.621	3.561
	(0.068)	(0.069)	(2.6)	[0.431]	[0.000]
Texas	0.086	0.082	-5.6	4.837	7.244
	(0.046)	(0.047)	(86.2)	[0.028]	[0.000]
W N Central Div	0.145	0.131	-7.4	4.259	10.384
	(0.052)	(0.053)	(12.1)	[0.039]	[0.000]
WV, VA, DC, MD, DE	0.241	0.228	-5.6	19.594	16.208
	(0.038)	(0.038)	(2.6)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.

Table F27: Return to Adv. Deg. STEM majors in Related occupation, by state (uncorrected and corrected)

Location	Uncorrected STEM Return	Corrected STEM Return	Percentage Change	χ^2 Test for Difference	F Test for Correction Terms
California	0.229	0.196	-15.9	39.205	39.768
	(0.029)	(0.029)	(3.6)	[0.000]	[0.000]
E S Central Div	0.219	0.199	-10.6	3.760	6.541
	(0.037)	(0.038)	(4.4)	[0.052]	[0.000]
Florida	0.269	0.251	-7.6	9.282	9.504
	(0.041)	(0.042)	(3.8)	[0.002]	[0.000]
Illinois	0.144	0.104	-29.0	25.491	28.640
	(0.040)	(0.041)	(17.8)	[0.000]	[0.000]
Mountain Div	0.215	0.212	-1.8	0.737	6.660
	(0.036)	(0.036)	(2.8)	[0.391]	[0.000]
NC, SC, GA	0.230	0.213	-6.7	15.571	14.025
	(0.031)	(0.031)	(2.9)	[0.000]	[0.000]
New England	0.131	0.090	-35.9	99.800	34.959
	(0.030)	(0.030)	(11.1)	[0.000]	[0.000]
New Jersey and Penn.	0.267	0.201	-20.1	74.933	39.428
	(0.029)	(0.030)	(4.1)	[0.000]	[0.000]
New York	0.183	0.132	-26.4	18.933	47.313
	(0.030)	(0.032)	(8.4)	[0.000]	[0.000]
OH, IN, MI, WI	0.203	0.172	-14.8	21.761	42.131
	(0.028)	(0.029)	(4.1)	[0.000]	[0.000]
OK, AR, LA	0.150	0.129	-15.5	3.002	4.082
	(0.054)	(0.055)	(84.3)	[0.083]	[0.000]
OR, WA, AK, HI	0.154	0.149	-6.2	2.698	10.790
	(0.048)	(0.048)	(5.8)	[0.100]	[0.000]
Texas	0.123	0.093	-26.6	33.742	18.653
	(0.035)	(0.035)	(13.1)	[0.000]	[0.000]
W N Central Div	0.191	0.190	-4.0	0.034	17.671
	(0.032)	(0.034)	(5.6)	[0.853]	[0.000]
WV, VA, DC, MD, DE	0.232	0.213	-7.2	13.604	20.536
	(0.029)	(0.029)	(2.7)	[0.000]	[0.000]

Notes: Bootstrapped standard errors (500 replications) below coefficients in parentheses. p-values below test statistics in brackets. See Appendix D for further details.