Major Complexity Index and College Skill Production [*†‡]

Xiaoxiao Li

Sebastian Linde

Hajime Shimao

April 15, 2021

Abstract

We propose an easily computable measure called the Major Complexity Index (MCI) that captures the latent skills taught in different majors. By applying the Method of Reflections to the major-to-occupation network, we construct a scalar measure of the relative complexity of majors. Our measure provides strong explanatory power of major average earnings and employment. Further evidence suggests that the MCI is strongly associated with advanced skills such as quantitative problem-solving, and the use of computing technology. We also provide a two-stage algorithm to partial out selection on observables which opens up possibilities of applying the complexity measure in various contexts.

Keywords: Major-to-Occupation Network; Skill Acquisition; Method of Reflections; Complexity.
JEL classifications: I2, J2

[*]We are grateful to Timothy Bond, Clint Harris, Teresa Harrison, Christopher Kilby, Elif Kubilay, Michael Lovenheim, Kevin Mumford, Jeffrey Smith, Yang Song, Evan Totty, Muzhe Yang, and participants at Eastern Economic Association Annual Conference, Western Economic Association International, 2021 CSWEP CeMENT Mentoring workshop, and Midwest Economics Association Annual Conference for helpful comments.

[†]The authors declare no relevant or material financial interests that relate to the research described in this paper. This paper uses public use data files that are available for download at the National Science Foundation (see: `https://ncsesdata.nsf.gov/datadownload/`) for the main analysis. Secondary analysis uses confidential data from the National Survey of Student Engagement (NSSE). This data can be obtained by application. (see: `https://nsse.indiana.edu/nsse/index.html`).

[‡]Contact information: Xiaoxiao Li, Villanova University, xiaoxiao.s.li@villanova.edu; Sebastian Linde, Medical College of Wisconsin, slinde@mcw.edu; Hajime Shimao, Santa Fe Institute, hshimao@santafe.edu.

1 Introduction

The return to education varies widely across fields of study in college (Altonji et al., 2012, 2016).[1] This heterogeneity is partly due to the fact that different majors send students to differing sets of occupations. For example, students with a petroleum engineering degree can find high-paying jobs as petroleum engineers, which are not easily accessible to students from many other majors. From the perspective of human capital accumulation, this linkage between college majors and students' occupational outcomes largely reflects the match of multi-dimensional skills.[2] That is, through various college majors, students acquire different sets of skills, which are subsequently employed in diverse job tasks.

The natural question to ask, then, would be: which major equips students with the most applicable set of skills? There is considerable difficulty when striving to answer this question. First of all, the skill acquisition process through different college majors is unobservable and thus needs to be indirectly inferred. Moreover, it is challenging and potentially dangerous for researchers to define what constitutes a "skill". In the previous literature, skills are typically defined from our intuitions, in low dimensions, such as verbal and quantitative, or cognitive and non-cognitive (e.g., Kinsler and Pavan, 2015; Heckman et al., 2006), rather than suggested by data empirically.

In this paper, we take a drastically different approach. By adopting the so-called "building-block" model proposed by Hidalgo and Hausmann (2009) in the context of international trade, we create an indirect measure of the skill set acquired through college education for each major. The intuition of the "building-block" approach is rather straightforward. In the context of education, each skill can be viewed as a different type of building block, like a Lego piece of a different shape. Each occupation is a Lego model, which requires a unique combination of building blocks. In order for a college graduate to find a job within an occupation, she needs to obtain the required

[1] Altonji et al. (2012) shows that the wage gap between electrical engineering and general education majors is almost as large as that between college and high school graduates.

[2] As the skill-portfolio analysis literature (e.g. Silos and Smith, 2015) suggests, some skills are very specific to certain occupations while others are more generally applicable. Conversely, some skills are easily attainable from many majors (e.g. oral communication) while others are taught in specific majors (e.g. understanding of thermodynamics that is commonly trained in a petroleum engineering major).

set of building blocks to form such a model. Now, the role of college majors is clear in this analogy: Each major is a bucket of building blocks where students can pick up the required pieces from.

What we need is a measure that captures the number and variety of blocks available in each bucket (i.e. college major). One obvious candidate is the number of occupations a major sends its students to. For instance, petroleum engineering and education administration majors both send graduates to eight distinct occupations, which may seem to imply that they teach similarly complex skills. However, this simple counting method would miss important information on the difficulty and specificity of skills that can be acquired from different majors. For example, one major may send students to occupations that only require a minimal skill set students could acquire from a large number of majors. In contrast, students from another major may find jobs in occupations where students from other majors cannot easily get in. In such cases, we would think the latter equips students with a set of more valuable, non-substitutable skills.

To construct such a measure, we adopt the "Method of Reflections" technique introduced by Hidalgo and Hausmann (2009).[3] This method exploits the rich information embedded within a major-to-occupation flow network and recovers the latent structure of necessary building blocks by an iterative algorithm. This paper is, to our knowledge, the first attempt to apply the Method of Reflections in the labor and education context.

Specifically, by incorporating information from the entire bipartite major-to-occupation network, we are able to uncover the underlying tripartite network connecting college majors to the skills they produce, and occupations to the skills they require. We call this measure the Major Complexity Index (MCI). The MCI takes into account relevant neighboring information in the network so that majors with exactly the same spread (number of occupations a major sends students to) can yield very different complexity ranks. Using the example above, petroleum engineering and education administration majors are both linked to eight occupations: the former is tied with occupations that are rarely accessible (less "ubiquitous" in the terminology of Hidalgo and

[3]Hidalgo and Hausmann (2009) introduce this method to take the bipartite network between countries and their exported goods and measure the latent production capacities and technologies that countries possess in order to produce the basket of observed exports. After its first introduction, this Economics Complexity Index (ECI) has been extensively utilized in the international trade literature (Tacchella et al., 2012; Mealy et al., 2019).

Hausmann, 2009) to other majors and thus returns a high major complexity index rank (2/137); while the latter maps students to occupations that are linked to many other majors, each of which is not highly ranked within the MCI, and in turn yields a lower MCI rank (135/137). The underlying idea is that occupations that require a complex set of skills are linked to, on average, majors that teach complex skills, while majors that equip more complex skills can send students to occupations that are more demanding. Thus, the complexity index has to be constructed recursively.

We explore three versions of the MCI according to a: (1) Binary flow matrix; (2) Weighted flow matrix based on the distribution of students within a major-to-occupation network; and (3) Controlled flow matrix based on the average marginal effects from a multinomial logit framework that accounts for potential bias due to selection on observables. One caveat of using the Method of Reflections in education-labor contexts is selection bias. Our modified two-stage algorithm (Controlled MCI) allows us to remove selection on observables, and as such, it opens up possibilities of using the complexity measure in settings where selection bias would otherwise be a concern.

Our empirical analysis employs individual-level information, including college major choice and occupational outcome, from the National Survey of College Graduates (NSCG) data in 2003, 2010, and 2015, to construct a bipartite major-to-occupation network for each year and examine the relationship between the MCI measures and average earnings as well as employment differentials across college majors. Our results indicate that the MCI reveals important aspects of college majors that matter to earnings (especially in recent years) and employment of college graduates. For example, using NSCG 2015 data, 1 standard deviation increase in the Binary MCI raises salary by $\$12,821$ or 16.9%, and boosts employment by 1.78 percentage points. An interesting observation within our study is that the power of the MCI to explain across-major wage differentials has increased considerably between the early 2000s and 2015, and decreased for employment. It is plausible to believe that as the structure of the economy transitions to a more technology-based era, the major complexity provides more insights into the earning differentials rather than the employment margin. This is consistent with recent work by Acemoglu and Autor (2011) which suggests that the rapid diffusion of new technologies may distort the earning distribution in a way

that benefits high-skilled workers.

In order to better understand what the MCI captures, we combine our major-to-occupation flow data from the NSCG with major level characteristics from the National Survey of Student Engagement (NSSE). Our results suggest that high MCI majors tend to have students with better pre-college academic qualifications, e.g. higher SAT scores in all three dimensions (mathematics, verbal reasoning, and writing ability), especially through a strong positive correlation with SAT math. In addition, high MCI majors are more intense and demanding in terms of time spent preparing for class, completing problem sets, and working on longer written assignments. More interestingly, in terms of knowledge and skills acquired through college education, students within high MCI majors tend to report further development of quantitative and practical problem solving, and the use of computing and information technology, but not in terms of basic skills such as written and spoken communication (which presumably are developed mainly through primary and secondary education). There are surprisingly few easily-computable quantitative descriptions of college majors, with limited examples such as major average SAT scores. Our comprehensive measure of major complexity is simple to compute with a minimum data requirement. It not only facilitates us to better understand the unobserved skill production process through college majors, but also provides a convenient and informative reference for both prospective students in choosing college majors, and education administrators in strategic planning, especially in resource-constrained environments.

Overall, our work contributes to a rich literature on the skill formation in college. This literature has been confronted with important challenges that we believe our approach is able to circumvent. Firstly, skills are presumably high dimensional. In the seminal paper, Cunha and Heckman (2007) identify two dimensions of skills, cognitive and non-cognitive, by means of a factor model. These two dimensions may likely be the most important distinction of skills, especially for early childhood education. In the context of college education, however, more fine-grained skill categories are necessary. For instance, leadership, a specific skill within the non-cognitive domain, is documented to have predictive power of potential earnings (Kuhn and Weinberger, 2005). A recent paper

by Deming (2017) highlights the importance of social skills in the labor market. Note, other intuitive measures of skills, such as quantitative versus verbal, are too coarse for the same reason. Secondly, the knowledge and skills acquired in college are particularly of interest by employers, and yet it is extremely difficult to measure the skills obtained by students that are factored in hiring decisions. On the occupation side, many studies have analyzed the skills required by different occupations (e.g. Graetz and Michaels, 2018). However, it is not an easy task to quantify the same skills acquired in college majors. For instance, how do we measure the programming skills that students can obtain from an economics major on average? Furthermore, the demand for skills constantly evolves over time as technology exponentially advances, and it is challenging yet critical to scrutinize the responses to such changes by college majors. Finally, one important, but under-explored aspect is the complementarity among skills (Cunha et al., 2006). Often a job task requires a combination of skills. For example, to be a financial engineer, one not only needs to be skilled in financial econometrics and programming but also management and communication that complement the technical background and help to improve job performance. Even if we fully observe the skill production process, with high dimensionality, it rapidly becomes impossible to estimate the complementary effects of every combination of fine-grained skill categories.

In summary, all of these challenges make our approach particularly appealing. Our proposed method does not intend to solve these issues, but rather circumvents them. To see this, it is important to note that the "building block" model fully captures the high-dimensional and combinatorial nature of skills, while the computation of the MCI avoids explicit estimation of the acquired skills and directly uncover the relative value of majors. That is, by exploiting the match of students between college majors and occupations, the MCI infers the latent skills taught in different majors and that are required by different occupations.

The rest of the paper is organized as follows: Section 2 introduces the Method of Reflections in the context of major-to-occupation network. Section 3 details the data sources. Section 4 presents our empirical results, and Section 5 discusses the limitations and important implications from both students and policymakers' perspectives.

2 Methods

Suppose we have $\mathscr{M}$ college majors and $\mathscr{O}$ occupations. Let N be a $\mathscr{M} \times \mathscr{O}$ flow matrix that represents the majors to occupations network. We first consider the flow matrix to be a binary matrix where $N_{m,o} = 1$ if and only if there exists at least one student who graduates from major m and finds a job in occupation o, and $N_{m,o} = 0$ otherwise. Using the Method of Reflections introduced by Hidalgo and Hausmann (2009), we iteratively calculate the value for each major and occupation according to equation (1) and (2), respectively:

$$k_{m,b} = \frac{1}{k_{m,0}} \sum_{o \in \mathscr{O}} N_{m,o} k_{o,b-1} \tag{1}$$

$$k_{o,b} = \frac{1}{k_{o,0}} \sum_{m \in \mathscr{M}} N_{m,o} k_{m,b-1} \tag{2}$$

where $b = 1, ..., B$ refers to the number of iteration and the initial values ($b = 0$) are the raw counts from the flow matrix:

$$k_{m,0} = \sum_{o \in \mathscr{O}} N_{m,o}$$

$$k_{o,0} = \sum_{m \in \mathscr{M}} N_{m,o}$$

Intuitively, $k_{m,0}$ represents the "spread" of major m in terms of the total number of occupations students can get in after graduating from major m. Similarly, $k_{o,0}$ represents the "specificity" of occupation o in terms of the count of majors that can place students in occupation o.[4] On the major side, a larger spread indicates higher complexity (as such a major can lead students to more occupations), while on the occupation side, smaller specificity implies higher complexity (as such an occupation accepts students only from limited majors). Upon convergence, the even iteration on the major side produces what we call the Binary Major-Complexity-Index (MCI).[5] The following example illustrates the Method of Reflections in the context of major-to-occupation flow network.

[4]The term "spread of a major" and "specificity of an occupation" correspond to "diversification of a country" and "ubiquity of a product" in Hidalgo and Hausmann (2009). Either way, they represent the number of links directly connected to a node.

[5]A similar index can be computed on the occupation side which is beyond the focus of this paper.

Example of Binary MCI

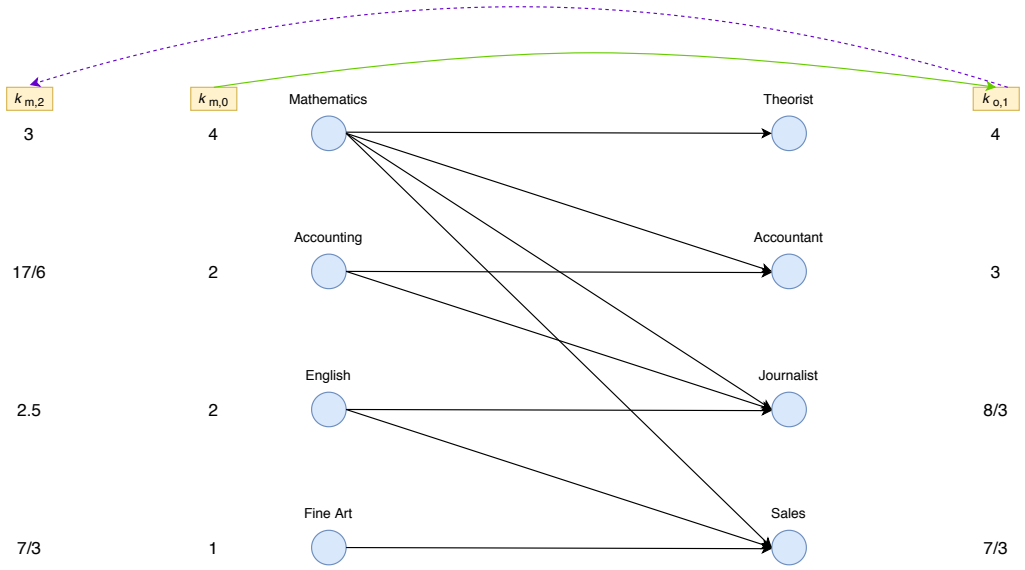

Figure 1: Bipartite Major-to-Occupation Network Example

Suppose there are four majors and four occupations, linked in a network as shown in Figure 1. For $m = Mathematics$,

$$k_{m,0} = 4,$$

indicating that students with a math degree find jobs in four distinct occupations (i.e. the spread of Mathematics is four). This is the naive complexity index at $b = 0$ where Mathematics is ranked in first place, followed by Accounting, English, and lastly Fine Art. The intuition is that the more occupations a major directs students to, the more complex the major is in terms of skills taught. However, this naive measure does not account for the accessibility of occupations. Using the example in Figure 1, the spread of Accounting and English majors are both two, with a common occupation outlet (Journalist). The difference is that Accounting can lead students to become Accountants, which is less accessible to students from other majors, when compared to Sales that is the alternative occupational outcome to an English major.

To incorporate this useful information within the network, we exploit the iterative procedure called the Method of Reflections. We first use the spread $k_{m,0}$ to calculate the complexity of each

occupation, $k_{o,1}$, in the first iteration ($b = 1$). As an example, for $o = Accountant$,

$$\begin{aligned} k_{o,1} &= \frac{1}{k_{o,0}} \sum_{m \in \mathscr{M}} N_{m,o} k_{m,0} \\ &= \frac{1}{2}(1 \times 4 + 1 \times 2 + 0 \times 2 + 0 \times 1) = \frac{4+2}{2} = 3. \end{aligned}$$

That is, there are two majors, Mathematics and Accounting, that place students as Accountants ($k_{o,0} = 2$), and the average spread ($k_{m,0}$) of these two majors is calculated to be 3. The same calculation can be carried out for all occupations $o \in \mathscr{O}$, as shown in Figure 1, $k_{o,1}$ column. Again, using the example in this network, Journalist and Sales are both connected to three majors (the occupation specificity $k_{o,0} = 3$), but the spread of majors that are linked to Journalist is, on average, greater than that for Sales. As a consequence, Journalist is ranked higher than Sales in this iteration. Importantly, majors are connected to one another through the occupation nodes within a network, and this intermediate step $k_{o,1}$ is essential to reflect such connections.

We then iterate back ($b = 2$) to update the complexity measure of each major $k_{m,2}$ using those values obtained in $k_{o,1}$. See the solid and dashed arrows on the top of Figure 1 for a graphic illustration of the iteration procedure. Back to the example of $m = Mathematics$,

$$\begin{aligned} k_{m,2} &= \frac{1}{k_{m,0}} \sum_{o \in \mathscr{O}} N_{m,o} k_{o,1} \\ &= \frac{1}{4}\left(4 + 3 + \frac{8}{3} + \frac{7}{3}\right) = 3, \end{aligned}$$

where $N_{m,o} = 1$ for all $o \in \mathscr{O}$ since the Mathematics major is linked to all four occupations, and thus is omitted from the calculation above. Intuitively, the average score of $k_{o,1}$ is three for Mathematics. Recall, $k_{o,1}$ captures the average spread of majors that are connected to an occupation. Following this logic, $k_{m,2}$ is basically the average of average spread of majors that are connected to one another through common occupational outcomes. What's embedded in this averaging process is the adjustment of major spread by occupation specificity. As shown in Figure 1, the complexity index of Accounting is updated from $k_{m,0} = 2$ to $k_{m,2} = 17/6 = 2.8\dot{3}$, while the adjustment for the

English major is only from $k_{m,0} = 2$ to $k_{m,2} = 2.5$ in this iteration. As a result, Accounting is now ranked higher than the English major, precisely for the aforementioned reason that Accounting can send students to occupations that are difficult to get into (i.e. Accountant in this example), and in turn it is connected to other majors that are relatively more complex in terms of skills taught.

While we stop the illustration at $b = 2$, we iterate on this procedure until its full convergence. The complexity index on the occupation side, $k_{o,b}$ for any odd number b, takes the average of the complexity scores of all majors linked to this occupation from the previous iteration, $k_{m,b-1}$. They are then used to compute the complexity index on the major side in the next iteration, $k_{m,b+1}$, which is the average of $k_{o,b}$ for all occupations linked to major m. The underlying idea is that occupations that require a complex set of skills are linked to, on average, majors that teach complex skills, while majors that equip more complex skills can send students to occupations that are more demanding. Thus, the complexity index has to be constructed recursively. Upon convergence, we obtain the Binary Major-Complexity-Index (MCI) on the major side, which is a specificity-adjusted spread taking into account relevant neighboring majors that map students into the same occupations. The intuition, which is further elaborated in Section 4.1, is that by utilizing the major-to-occupation bipartite network, we are able to shed light into the underlying tripartite network connecting college majors to the skills they produce, and occupations to the skills they require.[6]

Weighted MCI

The MCI calculation above utilizes a binary adjacency matrix. One could also construct the index using a weighted adjacency matrix instead. That is, if 100 students with a math degree find jobs in Theorist (10), Accountant (40), Journalist (20), and Sales (30), we then use the percentage of students from major m to each occupation as the proper weights (e.g., 0.1, 0.4, 0.2, 0.3) rather than equal weights (e.g. $\frac{1}{4}$) to calculate $k_{m,b}$. Similarly, if 100 students end up being Accountants,

[6]We elaborate on how a tripartite network of major-skill-occupation reduces down to a bipartite major-occupation network in Appendix A. The goal of the MCI is to infer the relative complexity of the skill set in each major based on the building block model (a tripartite network) from the information contained within a bipartite flow network. See Hidalgo and Hausmann (2009) and Hidalgo (2021) for more extensive discussions.

with 40 from Math and 60 from Accounting, we use the percentage of students from each major to occupation o as the proper weights (e.g., 0.4, 0.6) rather than equal weights (e.g. $\frac{1}{2}$) to calculate $k_{o,b}$. In this case, we call the converged even-iteration value on the major side the "Weighted MCI".

Controlled MCI

One caveat of using the Method of Reflections in the education-labor context is potential selection bias. That is, the apparent linkage between major m and occupation o could be partially due to factors other than skills match, such as gender preference. To account for selection on observable characteristics, we construct what we call the "Controlled MCI" using the multinomial choice framework. Specifically, we first regress the choice probability of each occupation on student characteristics, together with major dummies where a major dummy equals one if a student graduates in this major, and zero otherwise. That is, for each student i, let (m_i, o_i, x_i) be a tuple of his or her major choice, occupation choice, and other characteristics, such as age and gender. For each occupation $o \in \mathscr{O}$, we estimate the model:

$$Prob(o_i = o|m_i, x_i) = f\left(\sum_{m \in \mathscr{M}} \alpha_{m,o} \mathscr{I}(m_i = m) + \beta_o x_i\right),$$

where $\mathscr{I}(.)$ is an indicator function. We first conduct the estimation through a multinomial logit model[7], and then construct the adjacent matrix such that the (m,o)th element represents the average marginal effect of graduating with a major m on the likelihood of being in occupation o.[8] Similarly, we can construct our second weight matrix by regressing the choice probability of majors on occupation dummies, such that for each major $m \in \mathscr{M}$,

$$Prob(m_i = m|o_i, x_i) = f\left(\sum_{o \in \mathscr{O}} \alpha_{o,m} \mathscr{I}(o_i = o) + \beta_m x_i\right).$$

Intuitively, in the regression of occupation on majors, the average partial effects represent, on average, how likely a person is in this occupation from each major. Vice versa. Note that these

[7]Estimation is performed via the gradient descent algorithm on the maximum likelihood.

[8]The Method of Reflections requires all entries of the matrix to be non-negative. Thus we normalize the matrix by the min-max scaling.

regressions do not intend to infer causality between occupations and majors. The purpose is to construct weight matrices that capture the joint distribution of major and occupation as before while controlling for observables. Once we obtain the weight matrices, we can implement the Method of Reflections to construct the Controlled MCI.[9,10] This is done exactly as in the Weighted MCI approach, however, we now replace the matrix $N_{m,o}$ within equation (1) with controlled major matrix, and the matrix $N_{m,o}$ within equation (2) with the controlled occupation matrix.

3 Data

Our main data is sourced from the National Survey of College Graduates (NSCG) administered by the National Science Foundation. We use three cross-sectional datasets from the 2003, 2010, and 2015 surveys for our main analysis. One strength of the NSCG data is that it provides individual-level information on schooling (including major choice) along with occupational history, both of which are critical to construct a bipartite major-to-occupation network. We create three major-to-occupation networks, one for each survey year, with the following two data restrictions: First, in order to concentrate our analysis on the occupational placement of more recent graduates, we restrict our dataset to contain only individuals below the age of 40. Second, in order to minimize noise from spurious major-to-occupation linkages, we only keep majors and occupations with a minimum of at least 5 individuals in them.[11]

Table 1 provides descriptive statistics for our main outcome variables—major average salary

[9] Although, due to data limitation, we are only able to control gender and age in the individual regressions discussed above, our purpose here is to introduce this generalized procedure so that practitioners with better data availability (e.g. SAT scores, high school GPA, family background and resources, etc.) can utilize it to compute a robust major index.

[10] To identify the model, it is required that variables have sufficient variation in every major and every occupation. For instance, Geological Engineers occupation in 2003 only contain males (12 students) and Pre-school Teacher Education major in 2015 only contain females (42 students), and therefore are dropped from the analysis. For the same reason, we exclude ethnicity as a control variable as it results in omission of many majors and occupations.

[11] After imposing these restrictions, we have 27,852 observations in 2003, 24,315 observations in 2010, and 38,685 observations in 2015, that are used to construct major-to-occupation matrices. We are able to connect 140 majors to 92 occupations in 2003, 135 majors to 100 occupations in 2010, and 137 majors to 100 occupations in 2015. The number of majors and occupations are not restricted to be the same across years since they potentially reflect structural changes in the labor market. For instance, as technology reshaped the landscape in the 2000s, the national survey in 2010 and 2015 included trending occupations such as Computer Network Architect, Computer Programmer, Software Developers, Web Developer, etc. that were not captured in the 2003 survey.

in 2015 dollars (Salary) and major average employment rate (Employment Rate)—along with the major spread (Spread) and Major Complexity Index (MCI) obtained using the binary (MCI_B), weighted (MCI_W), and controlled (MCI_C) approaches. As shown in Panel A, there are a total of 137 majors in 2015 with an average major-level salary of $66,758 and employment rate of 91 percentage points. The mean major spread is about 37, implying that, on average, majors direct students to 37 distinct occupations. The major complexity indices are standardized to have within-sample mean of zero and variance of one.[12] Using the Binary MCI in 2015 as an illustration, the General Business major has a standardized index of zero, while Physics is one standard deviation above, and Elementary Teacher Education is one standard deviation below. Panel B and C display the major-level information of 2010 and 2003, respectively. The mean salary is slightly higher in 2010 compared to 2015; however, the variance, as well as the range (min and max), are smaller during the recovery period of the 2008 financial crisis. The lower mean employment rate tells a similar story. In contrast, the mild recession in 2001 does not have the same long lasting effects. The mean salary and employment rate are both higher in 2003 compared to 2010, where 76% of the majors, including Medical preparatory programs, Economics, and Accounting, experienced higher payments in real terms back in 2003, while other majors, such as Counseling Psychology, Statistics, and Petroleum Engineering enjoyed a wage premium surge in the late 2000s. Lastly, it's worth noting that the major spread is lower in 2010 compared to 2015, and even smaller in 2003.

To further understand what the Major Complexity Indices capture, we combine the 2015 NSCG data with a pooled cross-sectional dataset from the National Survey of Student Engagement (NSSE) for the years 2010-2011.[13] The latter contains rich student level data on the types of tasks and assignments performed across majors, such as what knowledge and skills are developed through college education as well as the hours spent on homework and papers. It also contains pre-college information, such as SAT scores. We are able to map 78 majors between the two datasets. Additional details pertaining to the matched dataset is provided in Appendix D.

[12]The indices are obtained after 250 iterations to ensure full convergence.

[13]NSSE surveys both freshmen and seniors in college. Our final NSSE sample of 2010 and 2011 data include 43% freshmen and 57% seniors who are most likely in the labor market by the time of 2015.

4 Results

The empirical results are organized into three subsections. We elaborate on the intuition of the MCI in Section 4.1, and present major-level regression results for both salary and employment in Section 4.2. Lastly, Section 4.3 provides an in-depth decomposition analysis of the MCI measure.

4.1 Spread versus MCI

As explained in the previous section, the Major Complexity Index (MCI) constructs the complexity measure of acquired skills by recursively considering the complexity level of other majors that map into the same occupations. By doing so, the MCI incorporates information from the whole bipartite major-to-occupation flow network and computes a comprehensive scalar measure of college majors that captures the latent skills taught in different majors and that are required by different occupations.

Figure 2 illustrates how the ranking of college majors change over iterations ($k_{m,0}$ to $k_{m,10}$) using the NSCG 2015 dataset. Here, we highlight two majors: Petroleum Engineering (yellow) and Education Administration (red). While both majors have the same spread ($k_{m,0}$) as their students are mapped into the same number of occupations (8 in both cases), the type of occupations that their students end up in differ considerably, and in turn we find that their major complexity indices ($k_{m,b}$ where $b =$ even number after convergence) are different. In the case of the Petroleum Engineering major, students are mapped into occupations that are rarely accessible to other majors such as petroleum and chemical engineers, and therefore we infer that the Petroleum Engineering major outputs students with skill sets that are hard to find elsewhere, and as such, it returns a high major complexity index rank (2/137). The reverse is true for the Education Administration major in which students are mapped into easily accessible occupations such as secondary school teaching and educational administration work, which in turn yields a lower MCI rank (135/137). The important takeaway here is that majors with exactly the same spread can yield very different complexity rankings due to the specific skill combination gained through each major which results in distinct

occupational outcomes.

4.2 Major Mean Wage and Employment Rate Analysis

While one could hypothesize that majors with greater spreads offer students a larger choice set of occupations and therefore generate higher option value, this does not appear to be the case as shown in Figure 3. This figure fits a linear regression between: (i) major mean salary against the spread ($k_{m,0}$) in the left-hand-side plot; and (ii) major mean salary against the Binary MCI in the right-hand-side plot. There is in fact no relationship between average salary and the spread, while there is a strong positive correlation between the average salary and the MCI. This highlights the intuition that variety of occupations by itself does not contain much information regarding the important skills acquired in each major. To discuss the value of a major, we need to consider the questions: Do the jobs available from this major require a complex skill combination that is not easily accessible from other majors, or do they demand only a minimal skill set that anyone could have? What comparative advantages does a major prepare its students in terms of skills and knowledge taught that are favorable and non-replaceable in the labor market?

The ordinary least squares results in Table 2 confirm our intuition above. Major spread is neither statistically nor economically significant in either level or log salary regressions, while the MCI measures are statistically significant at 1 percent level in explaining the earning differentials across college majors and substantially increase the explanatory power.[14] Using NSCG 2015 data in Panel A, one standard deviation increase in the Binary MCI raises major mean salary by $\$12,821$ in column (2) or 16.9% in column (6), and the effect is smaller using the Weighted MCI: $\$11,472$ in column (3) or 15.5% in column (7). The Controlled MCI paint a similar picture as shown in column (4) and (8) with even smaller estimates when we adjust the major-to-occupation flow matrix on the basis of age and gender, as described in Section 2. One standard deviation increase

[14]For instance, using NSCG 2015 data in Panel A, R^2 increases from 0.001 in column (1) to 0.522 in column (2) when the Binary MCI is controlled for. Here we note another interesting observation that the explanatory power of the MCI increases over iterations when we employ the Method of Reflections. As shown in Appendix C, R^2 increases considerably even between the 2nd and 10th iteration.

in the Controlled MCI raises salary by $\$10,739$ or 14.3%. Intuitively, the complexity of a major emerges from the number of marketable skills, the depth of each skill, and the interactions among those skills. And these skill sets provided by each major are highly correlated with students' potential earning outcomes. Similar patterns can be observed using NSCG 2010 and 2003 data in Panel B and C, respectively. Interestingly, the estimates are fairly similar for the 2003 and 2010 data (slightly higher in 2010 using the Weighted and Controlled MCI), but notably smaller in magnitude compared to the ones in 2015. The R^2 is also particularly large for the 2015 data.

Furthermore, the major complexity index is also statistically significant at the 1 percent level in explaining the employment rate differences across college majors, and it considerably increases the explanatory power, as shown in Table 3. Using NSCG 2015 data in Panel A, one standard deviation increase in the Binary MCI raises employment rate by 1.78 percentage points in column (2), or 1.97 percentage points in column (3) and 1.86 percentage points in column (4), using the Weighted and Controlled MCI, respectively.[15] Surprisingly, the estimated effect is larger in 2010 and the largest in 2003, and the R^2 is the highest in 2003. Together with results in the wage regressions, it is plausible to believe that the major complexity explains the employment and earning outcomes equally well. However, as the structure of the economy changes into more technology-based era, the major complexity provides more insights into the earning differentials rather than the employment margin. We provide further analysis of major ranking change over time in Appendix B, which lends supports to the argument that the observed MCI dynamics across time may reflect structural changes in the labor market.[16]

[15]Note, the effect is even larger if we instead use the residual employment rate where demographics such as age and gender are partialed out at the individual level. This set of results is available upon request.

[16]We also explore robustness exercises where major average SAT scores and other characteristics are controlled for in Appendix D. Even with a smaller sample size in the matched dataset, after controlling for students' academic qualifications (i.e. removing potential positive selection bias on preexisting abilities), majors with higher complexity scores still produce substantially higher average earnings. And in terms of the employment margin, controlling for additional major features results in larger estimates of the return to major complexity. See Table D.2 for more details.

4.3 Major Complexity Index Decomposition

Our preceding analyses suggest that the MCI reveals important aspects of college majors that matter to earnings (especially in recent years) and employment of college graduates. In order to better understand what the MCI captures, we combine the major-to-occupation flow data from the NSCG 2015 with major level characteristics from the National Survey of Student Engagement (NSSE) for the years 2010-2011. In total, we are able to match 78 majors between the two datasets.[17]

Table 4 provides the pairwise correlation between the Binary MCI and a number of major specific characteristics. First of all, in terms of students' academic preparation before coming to college, the MCI is positively correlated with students' performance in all three SAT measures (mathematics, verbal reasoning, and writing ability). Noticeably, the positive correlation between the MCI and SAT math score is particularly strong.

However, the MCI is not simply a repackaging of traditional measures, such as major average SAT scores.[18] Interestingly, when examining areas where students report that their current academic programs have helped them develop further knowledge and skills, we see that further development of quantitative problem solving, and the use of computing and information technology are strongly positively correlated with the MCI measure, with an estimated correlation of about 0.57 and 0.52, respectively. Applying theories or concepts to practical problems or in new situations also has a fairly strong positive correlation with the MCI. In contrast, there is a strong negative correlation between the MCI measure and the advancement of writing and speaking abilities in college. Taken together with the observed positive correlation with SAT verbal and writing scores, this suggests that higher MCI majors have students with high verbal and writing abilities (potentially developed within prior schooling), but who primarily report developing their ability to think critically and analytically, as well as the ability to analyze and solve quantitative and

[17] See Appendix D, Table D.1 for the summary statistics of the NSSE variables in the matched dataset.

[18] The major ranking based on the MCI is also not as simple as traditional categorizations such as STEM vs non-STEM majors, or rough area of study (e.g. Liberal Arts, Social Science, etc.). See Appendix B, Table B.1 - Table B.3 for the major ranking based on the Binary MCI for more details.

practical problems. Another observation worth noting is that, surprisingly, the MCI measure does not appear to be correlated with acquiring job or work-related knowledge and skills. Presumably, the robustness of the MCI in explaining wage and employment rate differentials across college majors is due to quantitative and analytical skills captured by the MCI rather than direct knowledge regarding the job content. We discuss this further in the next section.

In terms of time spent and efforts, we find that high MCI majors tend to have students who on average report spending longer hours preparing for classes, completing problem sets, and working on longer written assignments. If the time spent on studying can be viewed as a proxy for coursework intensity and difficulty, then this suggests that high MCI majors are more demanding on students and require them to invest more efforts into their schooling, which could generate higher payoffs.

5 Discussion

How are advanced skills formed through college education and is this skill production process responding to the changing nature of the economy? These are very hard questions to answer because skills are not directly observable, and some of them may not even be easily interpretable. In this paper, instead of explicitly modeling skill dimensionality, we take an alternative approach that computes a general measure of "complexity" for each major which reflects the skills taught in different majors and that are required by different occupations. Through the lens of this easily computable index, we can start to discuss these important questions.

Specifically, we apply the Method of Reflections introduced by Hidalgo and Hausmann (2009) to the major-to-occupation flow network and construct a scalar measure of the relative complexity in terms of skills taught in different majors. Our measure of complexity provides strong explanatory power in understanding average earning and employment variations across college majors. An interesting observation within our study is that the power of the MCI to explain across-major wage differentials has increased considerably between the early 2000s and 2015. An unexplored,

but potentially important, reason for this may lay with quantitative skills enjoying an increasingly larger wage premium in the labor market. Recent work by Acemoglu and Autor (2011) suggests that such a wage premium may derive from technological advancements that result in depressing wages in areas where machines are substitutes for laborers, while increasing wages in areas where machines are complements for workers. Our additional results provide further support along this avenue and suggest that the MCI strongly relates to advanced skills such as quantitative and practical problem solving, and the use of computing and information technology.

The major complexity indices exhibit rankings of college majors (see Appendix B, Table B.1 - Table B.3 for the major ranking based on the Binary MCI) that are naturally of interest to various stakeholders, including prospective students and their families in choosing college majors, as well as university administrators in charge of strategic planning of their schools. From students' perspective, it is essential to understand how the occupational outlook varies based on the choice of major. While it is well documented that expected earning is a key factor in choosing fields of study (Beffy et al., 2012; Wiswall and Zafar, 2015; Altonji et al., 2016), another equally important yet underexplored aspect is what occupations become available through the skills acquired in different college majors. Furthermore, as technology exponentially advances, skills valued by the labor market constantly evolve (Deming, 2017; Graetz and Michaels, 2018). The time trend of the major complexity ranking intuitively displays how each major is changing its relative position by modifying the bucket of "building blocks" as a response to the market demand for skills.

For administrators, the major complexity index presents a convenient and informative reference for their strategic planning. Evidently, colleges have been struggling to allocate resources across majors, particularly under a budget-constrained circumstance. For instance, University of Wisconsin at Stevens Point announced its elimination of 13 majors in 2018 to address "fiscal challenges" (Flaherty, 2018). Most recently, many universities are facing severe financial difficulties caused by the COVID-19 pandemic (e.g. Seltzer 2021), and are there-through forced to reallocate limited resources across majors. The major complexity index can facilitate this decision making process by providing information on which majors are preparing students with a more marketable combination

of skills. It is worth emphasizing that the MCI is easily computable with a minimum data requirement. Many universities have a center for career development which conducts post-graduation surveys or first-destination surveys. This allows administrators to apply our proposed method to their own major-to-occupation network to produce individualized major complexity ranking. They can also compare it against the national ranking of major complexity to better understand the comparative advantages of their own institution, and build their short- and long-term strategic plans accordingly.

One important caution in using the Method of Reflections is the assumption that major-to-occupation mapping is based on skills only. That is, the underlying model is that college majors produce skills, different occupations require different skill combinations, and a graduate finds a job if and only if the skills are matched. Conversely, if a linkage does not exist between a major and an occupation, it is only because the skills are not sufficient ("cannot get in") rather than nobody wants that job ("do not want to").[19] This concern is ameliorated to some extent by the fact that the National Survey of College Graduates (NSCG) surveys a large sample of college graduates in every wave. However, the potential selection bias remains as one caveat of applying this method to the labor environment. For instance, the apparent match between a major and an occupation could be partially based on factors other than skills, such as family resource or gender preference. Another contribution of this paper is to provide a two-stage algorithm (Controlled MCI) to partial out the selection on observables. Our algorithm opens up the possibilities of employing the complexity measure in various other contexts.[20]

Despite this limitation, the Method of Reflections and the complexity index are widely adopted in the international trade and development literature (See Hidalgo 2021 for a summary of applications in those fields) due to the minimum requirements of data and surprisingly strong explanatory power that this simple computation offers. Similarly, our generalized measure of MCI provides a useful tool for investigating difficult questions pertaining to the unobserved college skill production

[19]This critical assumption is acknowledged in the original Hidalgo and Hausmann (2009) study as well.

[20]If selection is on unobservables, then more structure is required.

process. As aforementioned, a major-to-occupation network can be easily constructed from many data sources and one can utilize the rich information contained in a network structure to extract the major complexity feature. Moreover, we can exploit the dynamics of such a network over time to explore any structural changes within college education as a response to the changing nature of the labor market.

Reference

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, Volume 4, pp. 1043–1171. Elsevier.

Altonji, J. G., P. Arcidiacono, and A. Maurel (2016). The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the Economics of Education*, Volume 5, pp. 305–396. Elsevier.

Altonji, J. G., E. Blom, and C. Meghir (2012). Heterogeneity in human capital investments: High school curriculum, college major, and careers. *Annu. Rev. Econ. 4*(1), 185–223.

Beffy, M., D. Fougere, and A. Maurel (2012). Choosing the field of study in postsecondary education: Do expected earnings matter? *Review of Economics and Statistics 94*(1), 334–347.

Cunha, F. and J. Heckman (2007). The technology of skill formation. *American Economic Review 97*(2), 31–47.

Cunha, F., J. J. Heckman, L. Lochner, and D. V. Masterov (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education 1*, 697–812.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics 132*(4), 1593–1640.

Flaherty, C. (2018). U wisconsin-stevens point to eliminate 13 majors. *Inside Higher Ed*.

Graetz, G. and G. Michaels (2018). Robots at work. *Review of Economics and Statistics 100*(5), 753–768.

Heckman, J. J., J. Stixrud, and S. Urzua (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics 24*(3), 411–482.

Hidalgo, C. A. (2021). Economic complexity theory and applications. *Nature Reviews Physics*, 1–22.

Hidalgo, C. A. and R. Hausmann (2009). The building blocks of economic complexity. *Proceedings of the national academy of sciences 106*(26), 10570–10575.

Kinsler, J. and R. Pavan (2015). The specificity of general human capital: Evidence from college major choice. *Journal of Labor Economics 33*(4), 933–972.

Kuhn, P. and C. Weinberger (2005). Leadership skills and wages. *Journal of Labor Economics 23*(3), 395–436.

Mealy, P., J. D. Farmer, and A. Teytelboym (2019). Interpreting economic complexity. *Science advances 5*(1), eaau1705.

Seltzer, R. (2021). N.J. university could cut 26% of full-time faculty amid budget woes. *Inside Higher Ed*.

Silos, P. and E. Smith (2015). Human capital portfolios. *Review of Economic Dynamics 18*(3), 635–652.

Tacchella, A., M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero (2012). A new metrics for countries' fitness and products' complexity. *Scientific reports 2*, 723.

Wiswall, M. and B. Zafar (2015). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies 82*(2), 791–824.

Figures and Tables

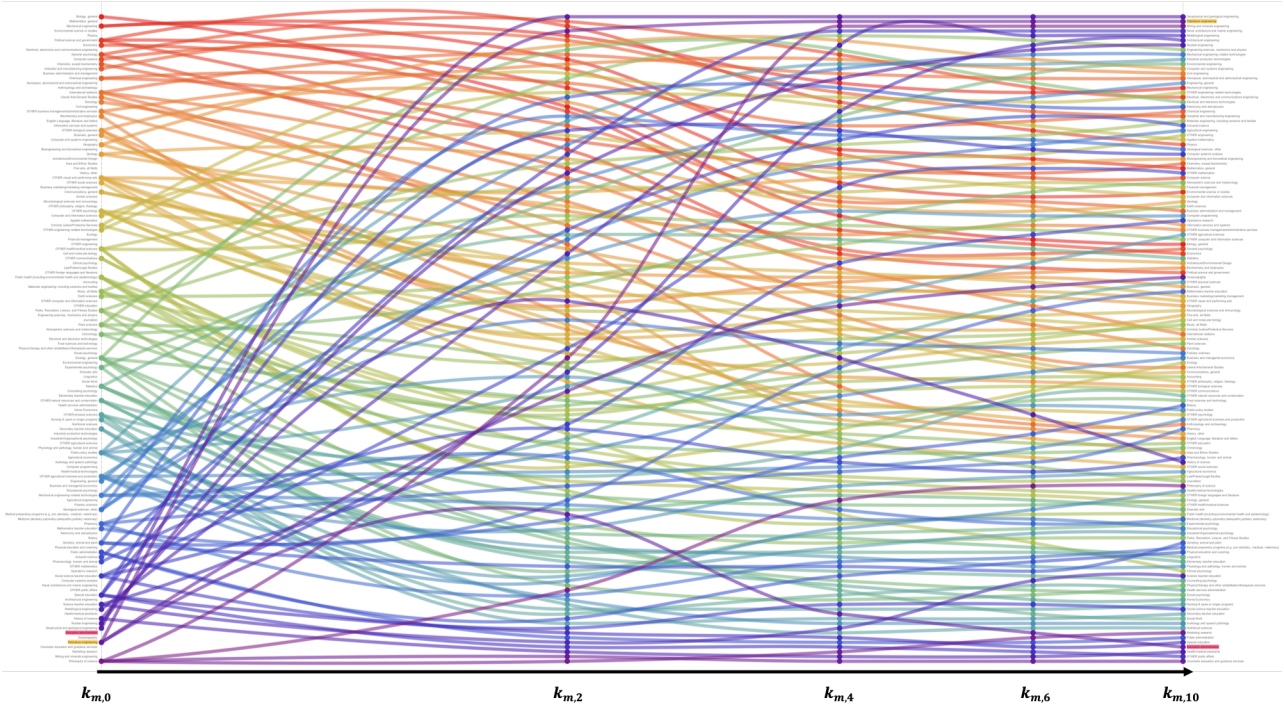

Figure 2: MCI ranking over iterations (Binary Adjacency Matrix Using NSCG 2015 Data). Iterations: 0,2,4,6, and 10 are reported. Two majors (Petroleum Engineering and Education Administration) that both map into 8 occupations are highlighted as yellow and red, respectively. Majors that are higher-up in the plot have higher MCI.

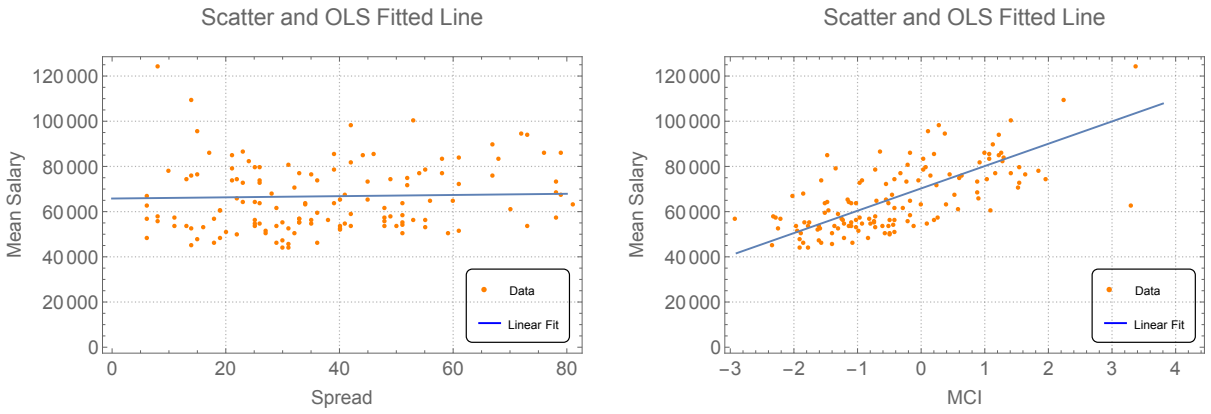

Figure 3: Scatter and Ordinary Least Squares Regression Lines. Using 2015 NSCG data, the left-hand-side (LHS) plot explores the relationship between major specific mean-salary and major-spread; while the right-hand-side (RHS) plot explores the relationship between major specific mean-salary and the Binary MCI.

Table 1: Summary statistics

	Mean	Std. Dev.	Min	Max	N
Panel A: NSCG 2015 Data					
Salary (2015$)	66,758.08	17,575.34	43,839.43	152,092.29	137
Employment Rate	91.01	4.62	75	100	137
Spread	36.69	19.19	6	81	137
MCI_B	0.00	1.00	-2.02	3.646	137
MCI_W	0.00	1.00	-1.78	3.286	137
MCI_C	0.00	1.00	-1.471	4.217	137
Panel B: NSCG 2010 Data					
Salary (2015$)	67,337.70	15,857.12	40,398.04	128,075.88	135
Employment Rate	88.36	5.41	69.23	100	135
Spread	31.01	17.16	5	81	135
MCI_B	0.00	1.00	-1.866	3.79	135
MCI_W	0.00	1.00	-2.001	3.406	135
MCI_C	0.00	1.00	-1.566	4.663	135
Panel C: NSCG 2003 Data					
Salary (2015$)	74,632.16	17,789.34	41,513.88	160,896.52	140
Employment Rate	89.14	5.52	66.67	100	140
Spread	29.96	16.19	3	71	140
MCI_B	0.00	1.00	-2.104	3.329	140
MCI_W	0.00	1.00	-2.425	3.858	140
MCI_C	0.00	1.00	-1.723	5.042	140

Note: Panel A reports the summary statistics for the NSCG 2015 data and major average statistics are computed using data from 38,685 students; Panel B for the NSCG 2010 data and major average statistics are computed using data from 24,315 students; and Panel C for the NSCG 2003 data and major average statistics are computed using data from 27,852 students. MCI measures are standardized to have sample mean 0 and standard deviation 1.

Table 2: Major Level Wage Regressions

	(1) Salary	(2) Salary	(3) Salary	(4) Salary	(5) ln(Salary)	(6) ln(Salary)	(7) ln(Salary)	(8) ln(Salary)
Panel A: NSCG 2015 Data								
Spread_2015	26.18	-71.33	-7.365	11.91	0.0010	-0.0003	0.0006	0.0008
	(93.45)	(73.99)	(71.99)	(73.35)	(0.0012)	(0.0009)	(0.0009)	(0.0009)
MCI_B_2015		12,821***				0.169***		
		(1,870)				(0.0191)		
MCI_W_2015			11,472***				0.155***	
			(1,928)				(0.0213)	
MCI_C_2015				10,739***				0.143***
				(2,282)				(0.0256)
Constant	65,798***	69,375***	67,028***	66,321***	11.04***	11.09***	11.06***	11.05***
	(4,331)	(3,314)	(3,373)	(3,398)	(0.0528)	(0.0386)	(0.0394)	(0.0404)
Observations	137	137	137	137	137	137	137	137
R-squared	0.001	0.522	0.426	0.374	0.007	0.517	0.441	0.382
Panel B: NSCG 2010 Data								
Spread_2010	67.02	-22.55	-0.624	22.895	0.00149	0.000179	0.000518	0.000848
	(87.01)	(78.75)	(72.49)	(75.77)	(0.00117)	(0.00103)	(0.000957)	(0.00100)
MCI_B_2010		7,801***				0.114***		
		(1,241)				(0.0165)		
MCI_W_2010			7,812***				0.112***	
			(1,434)				(0.0181)	
MCI_C_2010				6,775***				0.098***
				(1,250)				(0.0192)
Constant	65,260***	68,037***	67,357***	66,628***	11.05***	11.09***	11.08***	11.07***
	(3,431)	(3,164)	(3,001)	(3,110)	(0.0471)	(0.0420)	(0.0404)	(0.0420)
Observations	135	135	135	135	135	135	135	135
R-squared	0.005	0.238	0.243	0.186	0.013	0.254	0.250	0.205
Panel C: NSCG 2003 Data								
Spread_2003	115.9	115.6	111.6	128.2*	0.00172	0.00171*	0.00166*	0.00190*
	(80.00)	(71.27)	(70.87)	(74.82)	(0.00105)	(0.000923)	(0.000920)	(0.000971)
MCI_B_2003		8,489***				0.121***		
		(1,228)				(0.0161)		
MCI_W_2003			7,483***				0.108***	
			(1,532)				(0.0197)	
MCI_C_2003				6,265***				0.091***
				(1,663)				(0.022)
Constant	71,158***	71,170***	71,288***	70,790***	11.14***	11.14***	11.14***	11.13***
	(3,047)	(2,878)	(2,819)	(2,930)	(0.0395)	(0.0367)	(0.0357)	(0.0373)
Observations	140	140	140	140	140	140	140	140
R-squared	0.011	0.239	0.188	0.135	0.014	0.283	0.227	0.166

Note: Panel A reports results for the NSCG 2015 data; Panel B for the NSCG 2010 data; and Panel C for the NSCG 2003 data. The MCI measures are computed using 250 iterations. Robust standard errors are shown in parentheses. Significance is as follows: one-percent=***, five-percent=**, and ten-percent=*.

Table 3: Major Level Employment Rate Regressions

	(1) EmpRate	(2) EmpRate	(3) EmpRate	(4) EmpRate
Panel A: NSCG 2015 Data				
Spread_2015	0.0130	-0.0005	0.0072	0.0105
	(0.0218)	(0.0207)	(0.0194)	(0.0196)
MCI_B_2015		1.777***		
		(0.517)		
MCI_W_2015			1.970***	
			(0.373)	
MCI_C_2015				1.860***
				(0.404)
Constant	90.54***	91.03***	90.75***	90.62***
	(1.069)	(1.023)	(0.987)	(0.996)
Observations	137	137	137	137
R-squared	0.003	0.148	0.185	0.165
Panel B: NSCG 2010 Data				
Spread_2010	-0.0207	-0.0447	-0.0383	-0.0328
	(0.0282)	(0.0276)	(0.0254)	(0.0256)
MCI_B_2010		2.088***		
		(0.578)		
MCI_W_2010			2.027***	
			(0.411)	
MCI_C_2010				1.871***
				(0.396)
Constant	89.00***	89.74***	89.54***	89.37***
	(1.187)	(1.169)	(1.130)	(1.129)
Observations	135	135	135	135
R-squared	0.004	0.148	0.142	0.123
Panel C: NSCG 2003 Data				
Spread_2003	0.0270	0.0268	0.0256	0.0329
	(0.0303)	(0.0234)	(0.0255)	(0.0256)
MCI_B_2003		3.092***		
		(0.425)		
MCI_W_2003			2.476***	
			(0.404)	
MCI_C_2003				2.386***
				(0.422)
Constant	88.33***	88.34***	88.38***	88.19***
	(1.248)	(0.997)	(1.081)	(1.099)
Observations	140	140	140	140
R-squared	0.006	0.320	0.208	0.193

Note: Panel A reports results for the NSCG 2015 data; Panel B for the NSCG 2010 data; and Panel C for the NSCG 2003 data. MCI measures are computed using 250 iterations. Robust standard errors are shown in parentheses. Significance is as follows: one-percent=***, five-percent=**, and ten-percent=*.

Table 4: Pairwise Correlations Between the MCI and Major Specific Characteristics

Variable Description	MCI_B_2015
Standardized Test Scores	
SAT Verbal	0.36
SAT Mathematics	0.63
SAT Writing	0.29
Student Report - Developed Knowledge and Skills	
Writing clearly and effectively	-0.56
Speaking clearly and effectively	-0.59
Thinking critically and analytically	0.15
Analyzing quantitative problems	0.57
Using computing and information technology	0.52
Working effectively with others	-0.13
Learning effectively on your own	-0.14
Acquiring job or work-related knowledge and skills	0.05
Applying theories or concepts to practical problems or in new situations	0.46
Student Report - Time Spent	
Hours Spent Preparing for class	0.44
Amount of problem sets that take more than an hour to complete	0.63
Amount of problem sets that take less than an hour to complete	-0.17
Number of written papers or reports: 20 pages or more	0.14
Number of written papers or reports: between 5 and 19 pages	-0.24
Number of written papers or reports: fewer than 5 pages	-0.44

Note: Based on variation across 78 majors that are mapped between 2015 NSCG data and NSSE data for the years 2010-2011.

Appendix A Building Block Model and Flow Network

In this section, we briefly explain the intuition of the building block model in the context of major-to-occupation flow network. For more detailed discussion, see Hidalgo and Hausmann (2009).

Using the same example in Section 2 (Figure 1), there are four majors and four occupations. Now suppose the matching of students between majors and occupations are based on four latent, unobservable skils. On the left-hand-side of Figure A.1, a link between a major and a skill indicates that this major equips students with this skill. Conversely, a link between a skill and an occupation represents that this skill is required by that occupation. For example, students are required to obtain Skill 1 and 2 to be theorists. Since Skill 1 is only acquired in the Mathematics major, only students with a Mathematics degree can become theorists. In the similar logic, English and Fine Art majors cannot prepare students to be Accountants since Skill 2 is not taught in those majors.

Following this process, the tripartite network of major-skill-occupation reduces down to the bipartite major-occupation network on the right-hand-side. The goal of the MCI is to infer the relative complexity of the skill set in each major based on the building block model (left-hand-side figure) from the information contained within the flow network (right-hand-side figure).

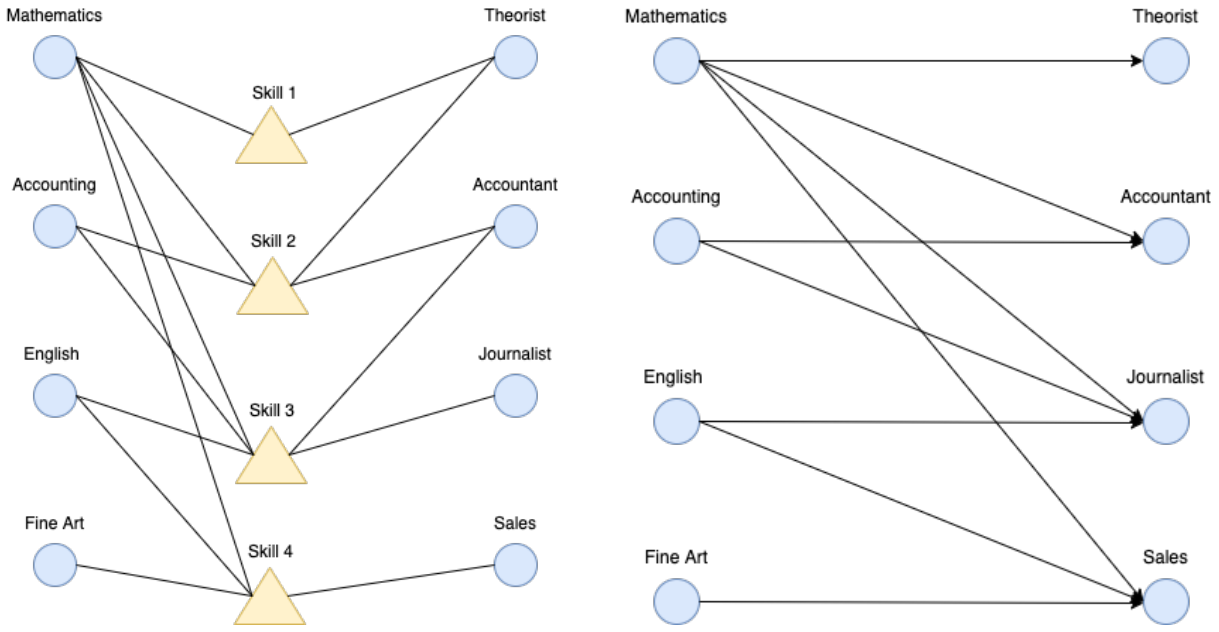

Figure A.1: Illustration of the Building Block Model.

Appendix B Major Complexity Index Rankings Over Time

Another interesting analysis based on our method is how the MCI ranking changes over time. Figure B.1 presents the dynamics of the major ranking using the Binary MCI over the 2003 to 2015 period. It reveals that while there is considerable consistency to many of the major rankings, some majors have experienced shifts across this time period.[21] For instance, Actuarial Science was ranked 96th in 2003, but rose to 37th in 2010, and further to a rank of 24th in 2015. It is interesting that many majors change their rankings non-monotonically. Given that the MCI explains the earning differences in all periods, we infer that these dynamics may reflect structural change of labor market conditions.

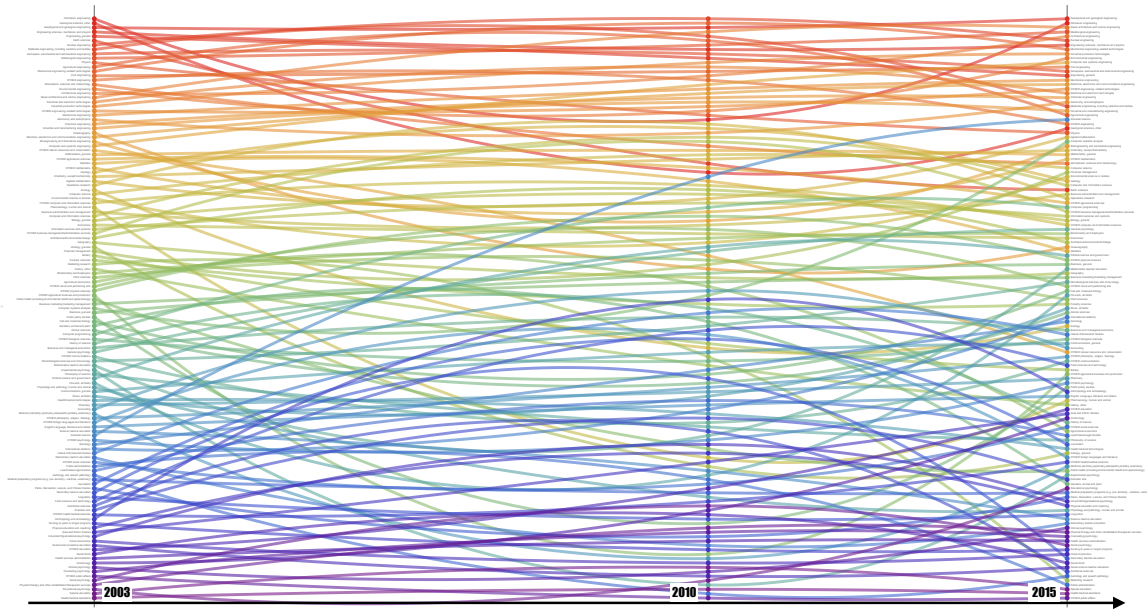

Figure B.1: Major Complexity Index (using the Binary MCI) Ranking Over Time (Years: 2003; 2010; 2015). Majors that are higher-up in the plot have higher MCI.

[21] The year-on-year consistency of the MCI ranking is qualitatively seen via the consistency of color-coding within Figure B.1 (i.e. red tends to stay on top, followed by orange, green, blue, and purple).

Table B.1: Major Ranking

Major Description	(1) STEM	(2) 2003	(3) 2010	(4) 2015
Geophysical and geological engineering	1	4	3	1
Petroleum engineering	1	2	24	2
Mining and minerals engineering	1	1		3
Naval architecture and marine engineering	1	20	4	4
Metallurgical engineering	1	11	7	5
Architectural engineering	1	19	14	6
Nuclear engineering	1	8	1	7
Engineering sciences, mechanics and physics	1	5	8	8
Mechanical engineering-related technologies	1	14	2	9
Industrial production technologies	1	22	11	10
Environmental engineering	1	18	22	11
Computer and systems engineering	1	31	16	12
Civil engineering	1	15	13	13
Aerospace, aeronautical and astronautical engineering	1	10	9	14
Engineering, general	1	6	10	15
Mechanical engineering	1	24	17	16
Electrical, electronics and communications engineering	1	29	18	17
OTHER engineering-related technologies	1	23	21	18
Electrical and electronic technologies	1	21	12	19
Chemical engineering	1	26	19	20
Astronomy and astrophysics	1	25	27	21
Materials engineering, including ceramics and textiles	1	9	5	22
Industrial and manufacturing engineering	1	27	20	23
Agricultural engineering	1	13	6	24
Actuarial science	1	100	37	25
OTHER engineering	1	16	15	26
Geological sciences, other	1	3	36	27
Physics	1	12	23	28
Applied mathematics	1	40	42	29
Computer systems analysis	1	71	64	30
Bioengineering and biomedical engineering	1	30	47	31
Chemistry, except biochemistry	1	39	33	32
Mathematics, general	1	34	31	33
OTHER mathematics	1	37	45	34
Atmospheric sciences and meteorology	1	17	30	35
Computer science	1	43	32	36
Financial management	0	57	57	37
Environmental science or studies	1	44	52	38
Geology	1	38	26	39
Computer and information sciences	1	48	28	40
Earth sciences	1	7	34	41
Business administration and management	0	47	48	42
Operations research	1	41	25	43
OTHER agricultural sciences	1	35	35	44
Computer programming	1	77	44	45
OTHER business management/administrative services	0	53	29	46
Information services and systems	1	52	43	47
Biology, general	1	50	41	48
OTHER computer and information sciences	1	45	39	49
General psychology	0	81	54	50

Table B.2: Major Ranking

Major Description	(1) STEM	(2) 2003	(3) 2010	(4) 2015
Biochemistry and biophysics	1	63	51	51
Economics	0	51	50	52
Architecture/Environmental Design	0	54	40	53
Oceanography	1	28	58	54
Statistics	1	36	38	55
Political science and government	0	87	53	56
OTHER physical sciences	1	67	46	57
Business, general	0	72	59	58
Mathematics teacher education	0	84	96	59
Geography	1	55	60	60
Business marketing/marketing management	0	70	61	61
Microbiological sciences and immunology	1	83	71	62
OTHER visual and performing arts	0	66	49	63
Cell and molecular biology	1	74	82	64
Fine arts, all fields	0	88	106	65
Plant sciences	1	64	56	66
Forestry sciences	1	59	66	67
Music, all fields	0	91	87	68
Animal sciences	1	76	84	69
International relations	0	103	90	71
Sociology	0	102	81	72
Ecology	1	42	76	73
Business and managerial economics	0	80	69	74
Liberal Arts/General Studies	0	104	68	75
OTHER biological sciences	1	78	79	76
Communications, general	0	90	77	77
Accounting	0	94	63	78
OTHER natural resources and conservation	1	32	55	79
OTHER philosophy, religion, theology	0	96	85	80
OTHER communications	0	82	93	81
Food sciences and technology	1	116	72	82
Botany	1	58	75	83
OTHER agricultural business and production	0	68	124	84
Pharmacy	1	93	86	85
OTHER psychology	1	101	97	86
Public policy studies	0	73	117	87
Anthropology and archaeology	0	120	65	88
English Language, literature and letters	0	98	89	89
Pharmacology, human and animal	1	46	102	90
History, other	0	62	80	91
OTHER education	0	127	109	92
Area and Ethnic Studies	0	123	103	93
Criminology	0	130	114	94
History of science	0	79	67	95
OTHER social sciences	0	106	108	96
Agricultural economics	0	65	62	97
Law/Prelaw/Legal Studies	0	108	88	98
Philosophy of science	0	86	94	99
Journalism	0	111	74	100

Table B.3: Major Ranking

Major Description	(1) STEM	(2) 2003	(3) 2010	(4) 2015
Health/medical technologies	1	92	91	101
Zoology, general	1	56	104	102
OTHER foreign languages and literature	0	97	73	103
OTHER health/medical sciences	1	119	112	104
Medicine (dentistry,optometry,osteopathic,podiatry,veterinary)	1	95	125	105
Public health (including environmental health and epidemiology)	1	69	100	106
Experimental psychology	1	85	132	107
Dramatic arts	0	118	92	108
Genetics, animal and plant	1	75	70	109
Educational psychology	0	136	130	110
Medical preparatory programs (e.g. pre-dentistry,-medical,-veterinary)	1	110	110	111
Parks, Recreation, Leisure, and Fitness Studies	0	112	120	112
Industrial/Organizational psychology	0	124	99	113
Physical education and coaching	0	122	101	114
Physiology and pathology, human and animal	1	89	111	115
Linguistics	0	115	116	116
Science teacher education	0	99	105	117
Elementary teacher education	0	105	122	118
Clinical psychology	0	131	128	119
Physical therapy and other rehabilitation/therapeutic services	0	135	118	120
Counseling psychology	0	132	119	121
Health services administration	0	129	121	122
Social psychology	1	134	113	123
Nursing (4 years or longer program)	0	121	123	124
Home Economics	0	125	126	125
Secondary teacher education	0	114	83	126
Social Work	0	128	115	127
Social science teacher education	0	126	129	128
Nutritional sciences	1	117	95	129
Audiology and speech pathology	0	109	134	130
Marketing research	0	60	78	131
Public administration	0	107	107	132
Special education	0	138	127	133
Health/medical assistants	0	140	133	134
Education administration	0	113		135
OTHER public affairs	0	133	135	136
Counselor education and guidance services	0	139		137
Science, unclassified	1	33		
Data processing	1	49		
Computer teacher education	0	61		
Pre-school/kindergarten/early childhood teacher education	0	137	131	

Note: Ranking based on the Binary MCI using 2003, 2010, and 2015 NSCG data, sorted by the 2015 Binary MCI.

Appendix C Wage Regression Results Across MOR Iterations

In applying the Method of Reflections (MOR) to our wage regressions, we find that the explanatory power (measured by the R^2 value) of the MCI is increasing over iterations. For example, comparing column (2) and (6) in Table C.1, R-squared increases from 0.321 to 0.519 between the 2nd and 10th iteration, using the binary transition matrix and 2015 NSCG data.

Table C.1: Wage regression results across iterations

	(1) Salary	(2) Salary	(3) Salary	(4) Salary	(5) Salary	(6) Salary
Spread_B_2015	26.18					
	(93.45)					
MCI_B_2015_iter2		9,954***				
		(1,665)				
MCI_B_2015_iter4			12,303***			
			(1,650)			
MCI_B_2015_iter6				12,631***		
				(1,692)		
MCI_B_2015_iter8					12,666***	
					(1,734)	
MCI_B_2015_iter10						12,658***
						(1,754)
Constant	65,798***	66,758***	66,758***	66,758***	66,758***	66,758***
	(4,331)	(1,242)	(1,076)	(1,048)	(1,045)	(1,046)
Observations	137	137	137	137	137	137
R-squared	0.001	0.321	0.490	0.516	0.519	0.519

Note: Binary MCI using NSCG 2015 Data. Robust standard errors are shown in parentheses. Significance is as follows: one-percent=***, five-percent=**, and ten-percent=*.

Appendix D NSSE Data Descriptive and Robustness Checks

To verify the robustness of major-level regression results presented in Table 2 and 3, we conduct further analysis controlling various features of college majors, including pre-college student characteristics such as SAT scores. To this end, we combine 2015 NSCG data with data from the National Survey of Student Engagement (NSSE) for the years 2010-2011 and compute average SAT scores of students in each major as well as major characteristics surveyed from students, such as knowledge and skills developed through college education and hours spent on coursework. We use data from these years since our final NSSE sample of 2010 and 2011 data include 43% freshmen and 57% seniors who are most likely in the labor market by the time of 2015. Table D.1 summarizes the variables used for the 78 majors that we are able to map between the two datasets.

Table D.2 reports regression results where major-level features are controlled for. Comparing column (1) and (2) in Panel A, we see that while adding controls for average SAT scores reduces the impact of the MCI on mean salary, it still remains statistically and economically significant. One standard deviation increase in the Binary MCI raises salary by $\$6,505$ in column (2), which implies that after controlling for students' academic qualifications (i.e. removing potential positive selection bias on preexisting abilities), majors with higher complexity scores still produce substantially higher average earnings. Additional major characteristic controls in column (5) further reduce the MCI estimate down to $\$4,571$, although, it is important to note that, because development of advanced knowledge and skills is the central channel through which the MCI affects earning outcomes, controlling for these additional characteristics may be an over-control for our purpose. Turning to Panel B of Table D.2, it is interesting that controlling for additional major features results in larger estimates of return (in terms of employment) to major complexity. For instance, one standard deviation increase in the Binary MCI raises the employment rate by 1.77 percentage points in column (2), and 1.86 percentage points in column (5).

Importantly, even with a limited sample size, these results indicate the robustness of the MCI in explaining the wage and employment rate differentials across college majors.

Table D.1: NSSE Summary statistics

Variable Description	Mean	Std. Dev.
Standardized Test Scores		
SAT Verbal	549.245	43.238
SAT Mathematics	559.034	51.246
SAT Writing	538.587	47.226
Student Report - Developed Knowledge and Skills		
Writing clearly and effectively	3.086	0.166
Speaking clearly and effectively	2.951	0.16
Thinking critically and analytically	3.36	0.09
Analyzing quantitative problems	3.094	0.231
Using computing and information technology	3.135	0.177
Working effectively with others	3.139	0.128
Learning effectively on your own	3.043	0.065
Acquiring job or work-related knowledge and skills	2.994	0.171
Applying theories or concepts to practical problems or in new situations	3.223	0.11
Student Report - Time Spent		
Hours Spent Preparing for class	4.447	0.398
Amount of problem sets that take more than an hour to complete	2.713	0.296
Amount of problem sets that take less than an hour to complete	2.542	0.204
Number of written papers or reports of 20 pages or more	1.472	0.132
Number of written papers or reports between 5 and 19 pages	2.427	0.215
Number of written papers or reports of fewer than 5 pages	3.047	0.174
N	78	

Table D.2: Salary and Employment Rate Regressions controlling Major Characteristics

Panel A:	(1) Salary	(2) Salary	(3) Salary	(4) Salary	(5) Salary
MCI_B_2015	8,442***	6,505***	5,948***	6,993***	4,571***
	(1,051)	(1,374)	(2,004)	(1,400)	(1,520)
SAT Verbal		-134.1			-27.09
		(84.32)			(122.5)
SAT Mathematics		150.5***			236.4*
		(48.87)			(119.8)
SAT Writing		-63.25			-107.8
		(64.02)			(87.95)
Student Report - Developed Knowledge and Skills			Yes		Yes
Student Report - Time Spent				Yes	Yes
Constant	64,698***	88,242***	32,924	-2,067	-112,731
	(1,168)	(14,959)	(67,768)	(43,602)	(87,897)
Observations	78	78	78	78	78
R-squared	0.404	0.520	0.557	0.643	0.723
Panel B:	EmpRate	EmpRate	EmpRate	EmpRate	EmpRate
MCI_B_2015	1.494***	1.770***	2.085***	1.405**	1.857***
	(0.343)	(0.495)	(0.629)	(0.545)	(0.651)
SAT Verbal		0.0245			-0.0739
		(0.0334)			(0.0508)
SAT Mathematics		-0.0135			0.0601
		(0.0152)			(0.0432)
SAT Writing		-0.0155			-0.00488
		(0.0237)			(0.0265)
Student Report - Developed Knowledge and Skills			Yes		Yes
Student Report - Time Spent				Yes	Yes
Constant	91.13***	93.58***	67.65***	111.8***	82.94***
	(0.404)	(6.442)	(20.89)	(16.98)	(30.40)
Observations	78	78	78	78	78
R-squared	0.151	0.161	0.388	0.232	0.491

Note: Panel A and B present results obtained by regressing major level average salaries and employment rate, respectively, on the Binary MCI using 2015 NSCG data and the major specific characteristics shown in Table D.1. Robust standard errors are shown in parentheses. Significance is as follows: one-percent=***, five-percent=**, and ten-percent=*.